EXPERIMENTAL DESIGN FOR MEASURING THE INTRA- AND INTER-GROUP CONSISTENCY OF HUMAN JUDGMENT OF RELEVANCE

A THESIS

Presented to

The Faculty of the Graduate Division

by

John Marion Hoffman

In Partial Fulfillment

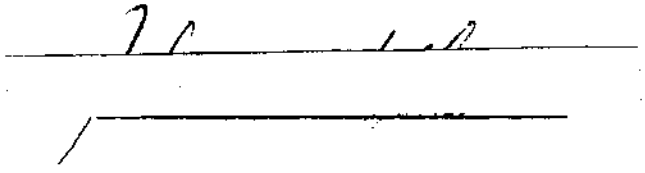
of the Requirements for the Degree

Master of Science in Information Science

Georgia Institute of Technology

August, 1965

In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institution shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.



EXPERIMENTAL DESIGN FOR MEASURING THE INTRA- AND INTER-GROUP CONSISTENCY OF HUMAN JUDGMENT OF RELEVANCE

Approved:

Date approved by Chairman: Aug. 13, 1965

ACKNOWLEDGMENTS

This study could not have been completed without the valuable assistance of the thesis advisor, Dr. Harrison M. Wadsworth and the members of the reading committee, Dr. Vladimir Slamecka and Mr. Dale L. Barker. The success of the pilot experiment was due mainly to the cooperation of Dr. Howard M. McMahon, the graduate students in the School of Aerospace Engineering at the Georgia Institute of Technology, and fellow students in the School of Information Science. Special thanks go to my wife, Susan, who spent many hours typing the preliminary drafts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

SUMMARY

The suspected variability of humans in judging the relevance of documents is one of the current problems confronting the development and improvement of document information and retrieval systems.

The purpose of this thesis was to design a method to investigate the variation, measured in terms of consistency, of relevance judgments between two groups of analysts and among the analysts within each group. To test the validity of the proposed design, a pilot experiment was conducted using two groups of analysts (subject experts and non-experts) and two question-document collections (machine retrieved and randomly selected). Analysts were instructed to mark each document relevant or not-relevant to the given question and to record the time required to make such relevance assessments. The responses were analyzed statistically to determine: (1) if the analysts within a group were consistent in their judgments of relevance; (2) if the two groups could be expected to make the same relevance judgment for the same question-document pair; (3) if one group was significantly more consistent in its assessments than the other; and (4) if the method of document selection had any effect on the consistency of relevance assessments. Expressions were formulated to show the degree of consistency exhibited by the analysts of each group.

The statistical procedures were of general utility under the experimental constraints. The data collected permitted, for the pilot experiment only, the following conclusions: (1) the analysts within the groups could consistently agree on the relevance of documents to questions;(2) the degree of consistency of the two groups did not

differ significantly; (3) the two groups did agree on the relevance of a particular document to a question; and (4) the method of document selection had a serious effect only on the consistency of the group of non-experts. The times required for the relevance judgments were in all cases lower for the expert group.

Analysis of the inconsistency among the analysts by means of relevance profiles indicated the probable need for relevance classes other than those of relevant and not-relevant.

CHAPTER I

BACKGROUND AND PURPOSE OF STUDY

The Problem of Relevance

What is relevance? The answer to this question is one of the current problems confronting the development and improvement of document information and retrieval systems. Despite a number of attempts to define relevance and characterize the process of relevance assessments, it is agreed that our understanding of it is shallow, and that more study is indicated.[1]

Most definitions of relevance offered so far have been operational definitions presented in connection with attempts to measure the efficiency of document storage and retrieval systems. In these instances, definitions of this term vary with different evaluation techniques or procedures. Thus in one instance, a relevant document is defined as that document from which the search question was made (the source document),[2] and in another, it is "that document which the questioner would like to have read before answering the question."[3]

---

[1] "Summary of the Study Conference in Evaluation of Document Searching Systems and Procedures," pp. 1-9.

[2] Cyril W. Cleverdon, "The Testing of Index Language Devices," ASLIB Proceedings, XV, p. 107.

[3] E. M. Fels, "Evaluation of the Performance of an Information Retrieval System by Modified Mooers Plan," American Documentation, XIV, p. 29.

In this connection, there has also been a recognition that relevance judging is not necessarily a dichotomous process, but that there may exist shades, or degrees, of relevance. Documents thus may fall into more than two groupings or categories. According to one view, there are five relevance categories: A document more useful than a "source document,"[4] a document as useful as the source document, a document of some interest, a document of no interest, and a false drop.[5] Another view defines three relevance categories -- crucial, relevant, and irrelevant.[6] Still another author requires only two such categories -- relevant and irrelevant.[7] Regardless of the number of relevance categories, this categorization is the prevalently accepted procedure for assessing the value of documents by those making the assessment.

If the relevance of a document to a search query is based on the value of the document to the assessor, what are the determinants in this mental process? The value of the document is determined according to one or more criteria (e.g., the understanding or interpretation of the search query, the relation of this document to another document, or the intensity of need or interest). In each case relevance is not a property of the document's content but a set of the assessor's criteria for value

---

[4]A source document is one from which the experimental question was formulated.

[5]Cleverdon, loc. cit.

[6]Harry Bornstein, "A Paradigm for a Retrieval Effectiveness Experiment," American Documentation, XIII, p. 255.

[7]A. Resnick and T. R. Savage, "The Consistency of Human Judgments of Relevance," American Documentation, XV, pp. 93-95.

judgments. The assessment of document relevance is then a process of matching the documents' content, as understood by the assessor, to a set of criteria which define the circumstances usually referred to as his "need," or "requirements," or "query."

If the set of criteria contains subjective elements, it is plausible to suspect that human judgment ("an operation of the mind involving comparison and discrimination"[8]) of relevance may vary with the assessor. This variation may be caused by value criteria such as educational background, experience in the subject field, personal motivation, and other physiological and psychological circumstances of the event. Thus it is proper to ask: What is the degree of agreement which may be expected of relevance judgments between certain types of individuals, and between certain categories of assessors? The answer or answers to this question have important practical implications.

Some Practical Applications of Relevance Assessments

Human judgment of the relevance of documents has been employed in most attempts at the evaluation of information retrieval systems; it is a standard function in the review of search outputs in both mechanized and manual information systems; it is a crucial function in the process of acquisition of materials in libraries; and so forth. As will be shown, little attention has been paid in these areas to the suspected variability of human assessments of relevance.

Various suggested methods of evaluating the relative efficiency

---

[8]Stanford L. Optner, Systems Analysis for Business and Industrial Problem Solving, p. 19.

of information retrieval systems, based on the measurement of the percentage of relevant material retrieved and the accompanying amount of irrelevant material, have employed human judgment. For example, Bornstein used "the user of the information and his judgment of the relevance of the information retrieved to specific questions,"[9] The Mooers Plan and a modification of it[10] utilized human judgment to arrive at a measure of system efficiency. Again, Swanson[11] has relied upon a group of physicists with post-doctoral experience in the specialties represented by the articles to make the required relevance judgments.

Typical of the conclusions reached is that of the Cranfield staff in an experiment at the English Electric Library at Whetstone.[12] It states that an experienced librarian was able to recognize documents relevant[13] to a set of questions in a "relatively strange subject field," and that he was as successful in searching for relevant documents as persons with practical experience in the subject field. Yet the design of the Whetstone experiment, conducted in a manner of earlier experiments

---

[9]Bornstein, op. cit., p. 254.

[10]Fels, loc. cit.

[11]Don R. Swanson, "Searching Natural Language Text by Computer," Science, CXXXII, pp. 1100-1101.

[12]Cyril W. Cleverdon, "ASLIB Cranfield Research Project, Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," pp. 61-75.

[13]The term "relevant document" in the Cranfield studies is the source document from which the question was made.

by the Cranfield staff,[14] did not admit of the variability of human assessments of relevance.

If it is true, as has been postulated,[15] that definitions and characteristics of relevance vary from one individual to another, the following questions arise with respect to studies of system efficiency: Had Bornstein selected several groups of three users each, would he have obtained similar judgments of relevance by each group? Or, in the Mooers Plan, would another representative sample of users (or another writer and umpire in the Modified Plan) have yielded the same efficiency rating? Again, would Swanson have reported similar results had he selected another panel of physicists? Were the librarians of the Whetstone study not guaranteed the existence of a relevant (source) document, would they still have performed as well as trained technicians, and identically with other librarians? All these studies leave these questions unanswered.

One of the proposed functions of an information system has been for personnel on the system staff to further analyze the collection of retrieved documents for irrelevant documents before passing the search results on to the inquirer. But what assurance is there to the inquirer

---

[14]Some 200 references were randomly selected from the indexed collection and were distributed among the Whetstone technical staff with instructions to prepare questions in such a manner that each question would be successfully answered by one and only one of the references (i.e., by that document from which the question emanated). Subsequently, searches were conducted until the source document was retrieved or until no further search programs could be devised.

[15]Donald J. Hillman, "The Notion of Relevance (1)," American Documentation, XV, p. 29.

that the documents rejected by the system analyst were not relevant? Is it not possible that some documents classed as irrelevant could be of some value for the user? A similar problem arises in the case of the librarian. Is it correct to assume that one librarian's criteria for classifying documents are the same as those of another librarian and those of information users? At present there seem to be no studies concerned with these problems.

The questions presented in the preceding paragraphs may be summarized as follows. (1) For the discussed information system evaluation techniques, are the efficiency ratings obtained independent of the selection of the user-experts who evaluate the relevance of search outputs? (2) Is the evaluation of the search results from both conventional and mechanized techniques independent of the personnel who does the screening? (3) Is the determination of a relevant document by a person independent of his educational background?

Previous Experiments to Evaluate Relevance Assessments by Humans

During his visit to the United States in 1961, Cleverdon was asked to consider that with regard to the determination of relevance in the Cranfield Project judgments about relevance of documents to a particular request vary (1) among users, and (2) with time for one user.[16]

Related to these suggestions is a series of experiments by Resnick, Savage, and Rath. Their first experiments[17] were conducted to determine which of four types of lexical indicators (i.e., document surrogates)

---

[16]John O'Connor, Journal of Documentation, XVII, pp. 259-260.

[17]G. J. Rath, A. Resnick, T. R. Savage, "Comparison of Four Types of Lexical Indicators of Content," American Documentation, XII, pp. 126-130.

could best be utilized by subjects to determine relevant from irrelevant documents. The results indicated that there was no major difference between the text and the abstracts. In a later experiment individuals were asked to determine relevance of documents to their work interests on the basis of titles and abstracts. It was indicated that there were no significant differences between the usefulness of titles and of abstracts for this purpose.[18]

Of primary interest to this study was the final experiment concerning the consistency of human judgment of relevance. Resnick and Savage[19] compared the relative inter- and intra-subject consistency of humans when judging the relevance of documents to their general interests on the basis of different lexical indicators.[20] The subjects making the judgments were divided into four groups, one for each indicator, and were instructed to judge the items of their group relevant, R, or irrelevant, I, to their interests. One month later the experiment was repeated using the same people and the same items with the additional instruction requiring each subject to recall which items he felt he had responded to in the same way in the first experimental session.

There were four possible categories of responses for any one item: R-R (i.e., the item was rated R on both the first and the second

---

[18] A. Resnick, "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents," Science, CXXXIV, pp. 1004-1005.

[19] Resnick and Savage, loc. cit.

[20] The functions of a lexical indicator of content were to determine if a document was relevant for a specific purpose and to obtain some information from the document without having to examine the entire text.

experimental sessions), R-I, I-R, and I-I. The frequency of the responses in the classes R-I and I-R was compared to the total frequency of responses of all four classes. In this experiment 10 per cent of the 46 response pairs were in the R-I, I-R classes and the remaining 90 per cent were in the classes R-R and I-I. Therefore, if a subject judged an item one way on the first session, then the indication was that he would make the same judgment 90 per cent of the time on his second judgment of the same lexical indicator. When the changes were analyzed statistically, the test indicated that the changes which occurred were not significantly different (i.e., there was no reason to believe the probability of an I-R pair of judgments was not equal to the probability of an R-I pair), except for abstracts.[21]

The data indicating the number of judgments recalled correctly on the second session were tabulated and tested by the Kruskal-Wallis One-Way Analysis of Variance test. It examined the hypothesis that each of the recall scores came from the same population or from identical populations with respect to averages. For the experiment the hypothesis could not be rejected; hence, there was indicated an inter-subject consistency. That is, the members of all groups were able to recall equally well previous judgments of relevance.

Another experiment was conducted by G. C. Barhydt at Western Reserve University to measure the effectiveness of relevance assessments based on non-user evaluation.[22] He indicated there was a "high

---

[21]There was no obvious reason for this occurrence.

[22]Gordon C. Barhydt, "A Comparison of Relevance Assessment by Three Types of Evaluator," Proceedings of the American Documentation Institute, I, pp. 383-385.

correspondence" of the test scores between the subject expert and the system specialist when judging the relevance of documents to questions. The experiment was conducted on a preliminary basis using only a limited number of question-document pairs and only one subject expert and one system expert. The author did state that future experiments would be conducted to further validate his conclusions.

Purpose of the Study

Variation of relevance judgments due to a time factor was investigated in the experiment by Resnick and Savage. The problem of varying relevance among individuals is still left to be investigated. To investigate this aspect and to attempt to provide answers to the questions presented earlier in this discussion, an experiment was designed and conducted to measure the inter- and intra-group consistency of the judgments of the relevancy of documents to specific questions by subject experts and non-experts. The goal sought was to design an experimental procedure of general utility, to develop valid statistical procedures for analyzing the data collected, and to perform a pilot experiment testing these procedures.

CHAPTER II

EXPERIMENTAL DESIGN

Definitions of Variables

The purpose of this study was to design and test an experiment to determine the intra- and inter-group consistency of relevance judgments. As indicated in Chapter I, there were many variables to be considered. A definition of relevance and the related relevance categories should be agreed upon. Were the retrieved documents to be compared to other retrieved documents, or were they to be evaluated on the basis of some predetermined criterion such as a search question? From what subject fields were the documents to be selected? What was to be the relationship of the document analysts to the subject fields? These variables and their definitions are given in the following paragraphs.

Since the purpose of this study was to examine intra- and inter-group consistency, the analysts within the groups, the groups themselves, and the search questions and retrieved documents must be considered as independent variables. For relevance judgments to have meaning, the subject field, an operational definition of relevance, and the established relevance categories remained constant throughout the experiment. Dependent upon the analyst was his set of value criteria and relation to the subject field.

A "relevant document" was that document which, in the opinion of the analysts, was of some interest with respect to the criterion of the search question. This definition was operational and was dependent upon

the analysts who established their own set of values. It was postulated that the operational understanding of relevance varied among analysts, and that such variations were reflected in an inconsistency of their relevance judgments.

The apparent existence of relevance "degrees" has been mentioned above. Since little work has been done to investigate human consistency of relevance evaluations, the simplest case when there were only two relevance classes -- relevant and not-relevant -- was considered for this study. If other degrees were necessary to better characterize the nature of relevance, a method based on the disagreement distribution among the analysts was presented to define these degrees.

It was necessary to consider samples from three populations: (1) Analysts, (2) Search questions, and (3) Retrieved documents. There were two groups of analysts. Group A, the experts, consisted of 14 graduate students enrolled in the School of Aerospace Engineering at the Georgia Institute of Technology.[23] Group B, the non-experts,[24] consisted of 14 graduate students of the School of Information Science. For identification purposes the members of each group were assigned three digit numbers (group A begins with "1" and group B a "2").

Twelve search questions in the subject fields of aerospace engineering were formulated by the faculty and students of the School of Aerospace Engineering. Nine of these questions were "answered" by the

---

[23]For a summary of definitions and symbols used in this discussion, see Appendix A.

[24]Those who were not engaged in the study or practice of aerospace engineering.

random selection of document abstracts (identified by document numbers prefixed with an "A" or "B") from Scientific and Technical Aerospace Reports (STAR). The other three questions were submitted to the Defense Documentation Center (DDC). The "answering" document abstracts for these questions were randomly selected from the output received from the mechanical retrieval system. It should be noted that the experiment did not require that the documents submitted to analysts for their judgment of relevance actually be retrieved answers to those questions; on the contrary, matching of randomly selected documents with arbitrary questions removed any inconsistency that could have been induced into the document collection by previous judgments of indexers. The effect of the two methods of document selection is examined later in the experiment.

Each set of numbered documents submitted in response to a particular question was presented individually to the analysts of the two groups. (Recall that the documents were matched to the questions by both random selection from STAR and actual machine search by DDC.) Each analyst independently and without additional reference to other individuals or reference material marked every document relevant (R) or not-relevant (I) to the appropriate problem question, and recorded for each question the time necessary to make the decisions for the related documents.

When all sets of questions and documents were reviewed individually and independently by all members of both groups, a response matrix was prepared for each group and each question. This matrix is of the form indicated by Figure 1.

	Documents								
	D_{1q}	D_{2q}	∘	∘	∘	D_{jq}	∘	∘	∘	D_{mq}
A_1	θ_{11q}	θ_{12q}	∘	∘	∘	θ_{1jq}	∘	∘	∘	θ_{1mq}
A_2	θ_{21q}	θ_{22q}	∘	∘	∘	θ_{2jq}	∘	∘	∘	θ_{2mq}
∘										
∘										
∘										
A_i	θ_{i1q}	θ_{i2q}	∘	∘	∘	θ_{ijq}	∘	∘	∘	θ_{imq}
∘										
∘										
∘										
A_N	θ_{N1q}	θ_{N2q}	∘	∘	∘	θ_{Njq}	∘	∘	∘	θ_{Nmq}
R_{jq}										
I_{jq}										
F_{Ajq}										
F_{Aq}										
F_A										

Figure 1. Response Matrix for Question q, Group A.

Hypotheses

Before continuing the discussion, it was necessary to define what was meant by the phrase "agreement in human judgment of relevance of documents to questions." Each member of the groups (A and B) analyzed the documents to determine if they were relevant (R) or not-relevant (I) to a particular question. If, for the jth document, the qth question and either group the number of R's, R_{jq}, was not equal to the number of I's, I_{jq}, (i.e., $R_{jq} \neq I_{jq}$), then for that group, question, and document agreement among the analysts had occurred. In other words, the analysts agreed that the document was either relevant or not relevant to the question.

For this experiment five hypotheses were tested. The first hypothesis consisted of one major hypothesis and a related hypothesis.

Hypothesis 1

There is an equally likely chance for the members of a group (group A or B) to agree or disagree on the relevancy of documents to questions. In other words, there is no agreement among the members of a group in their judgments of relevance of documents to questions.

Data for Hypothesis 1 was obtained by testing the following related hypothesis:

Hypothesis 1-a

There is an equally likely chance for a group to judge a document relevant (R) or not-relevant (I) to a question. That is, there is no agreement among the members of a group in their judgments of relevance of document j, question q. (For the remainder of this paper the notation D_{jq} is defined as "document j for question q.")

If the results allowed the rejection of Hypothesis 1, it was

concluded that the members of the group in question could consistently agree on the relevance of a document to a question. On the other hand, if the hypothesis was not rejected, then members of the group did not exhibit the ability to consistently come to any agreement. The degree to which the analysts agreed was measured by F_A -- the fraction of the members of group A who agreed on relevance judgments for all documents examined -- and by F_B -- the fraction scores for group B.

Hypothesis 2

The fraction of agreement scores, F_A and F_B, will be the same for the groups of experts and non-experts (i.e., $F_A = F_B$).

Two additional tests were necessary to complete the comparison of group A to group B. If the members of group A agreed that a document was relevant to a question while the members of group B agreed that it was not-relevant, identical scores, $F_A = F_B$, would have little meaning. Thus,

Hypothesis 3

There is an equally likely chance for group A to agree or disagree with group B on the relevancy of a particular document to a question. If p_{nc} is the probability that group A and group B made the same relevance judgment for a document and p_c is the probability that the two groups differ on relevance judgments for a document, then from Hypothesis 3, $p_{nc} = p_c = \frac{1}{2}$.

Hypothesis 4

For the judgments that differ from one group to another, the probability that a document will be judged relevant to a question by group A and not-relevant by group B (p_{RI}) is equal to the probability that a

document will be judged not-relevant to a question by group A and relevant by group B (p_{IR}).

Since several questions were answered by machine searches and the rest by the random matchings of documents to questions, it was worthwhile to consider the following hypothesis:

Hypothesis 5

The fraction of agreement scores will be the same for documents "retrieved" by random methods and mechanical search methods.

Data Analysis

The analysis of data collected employed the binomial test[25] for Hypothesis 1-a and an approximation of this test by the normal distribution for Hypothesis 1 and Hypothesis 3; for Hypothesis 2 and 5 the Kolmogorov-Smirnov Two-Sample Test[26] was used; and the McNemar Test for the significance of changes[27] was used to test Hypothesis 4. Hypothesis 1, 1-a, and 5 were tested for both groups A and B.

Test 1 (Hypothesis 1-a)

From the response matrix (see Figure 1) for question q, the number of R's and I's were counted for document D_{jq}. Since the analysts made their judgments without consulting any other references such as fellow students or reference works, each event of judgment was considered independent. There were two possible choices (R or I) for each event; hence, from Hypothesis 1-a, the probability of either choice was $p_R = p_I = 1/2$.

---

[25] Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences, pp. 36 - 42.

[26] Ibid., pp. 127 - 136.

[27] Ibid., pp. 63 - 67.

Since Hypothesis 1-a assumed $p_R = p_I = 1/2$, the probability of observing x responses of the same type is

$$p(x) = \binom{N}{x}\left(\frac{1}{2}\right)^N \tag{1.1}$$

where

$$\binom{N}{x} = \frac{N!}{x!(N - x)!} \tag{1.2}$$

and

$$N = I_{jq} + R_{jq} \tag{1.3}$$

If x is the smaller of the number of observed relevant and not-relevant responses for D_{jq}, then the probability that x is less than or equal to some integral constant, k, is

$$p(x \leq k) = \sum_{x=o}^{k}\binom{N}{x}\left(\frac{1}{2}\right)^N \tag{1.4}$$

For a significance level of α, if $p(x \leq k) < \alpha$. Hypothesis 1-a is to be rejected. In the pilot experiment a significance level of 0.05 was chosen permitting the scores $F_A jq$ and $F_B jq$ (see Equation 4.3) to be in the low 70 per cent range before rejecting Hypothesis 1-a. For the number of observations contained in the experiment and for $\alpha = 0.05$, it was observed that $p(x \leq 3) < 0.05$ permitting the rejection of Hypothesis 1-a when x was three or less.

Test 2 (Hypothesis 1)

Since the purpose of Hypothesis 1 was to test whether the analysts

of a group agreed on more documents than they disagreed, the hypothesis assumed that $p_o = p_1 = 1/2$. (p_o was the probability of not rejecting Hypothesis 1-a (i.e., disagreement) and p_1 was the probability of rejecting Hypothesis 1-a or agreement.) Since the number of documents analyzed in the pilot experiment, n, was much larger than 25 and from the assumption of Hypothesis 1, the normal distribution approximated the binomial. The expression for this approximation was:

$$z = \frac{x - np_o}{\sqrt{np_o p_1}} \qquad (2.1)$$

where x was the number of times Hypothesis 1-a was not rejected for the sample of documents analyzed (n = 202). For the significance level, $\alpha = 0.05$, the rejection region was all values of x such that $x \leq 88$. If this hypothesis was rejected, the alternative, $p_o < p_1$, indicated that the group was more likely to agree than disagree. In other words, the group was consistent in their judgments of relevance. Otherwise, if the hypothesis could not be rejected, the analysts of a group could not be expected to agree more than they disagree or that they were inconsistent in their relevance assessments.

Test 3 (Hypothesis 2)

Before testing Hypothesis 2, the percentage scores indicating the fraction of the members of a group making the same judgments must be determined. From Figure 1, when θ_{ijq} was R or I, φ_{ijq} was assigned 1 or 0 respectively. Thus

$$R_{jq} = \sum_{i=1}^{N} \varphi_{ijq}, \quad \begin{matrix} j = 1,2,\ldots,m_q \\ q = 1,2,\ldots,Q \end{matrix} \qquad (4.1)$$

$$I_{jq} = N - R_{jq}, \quad j = 1,2,\ldots,m_q \qquad q = 1,2,\ldots,Q \tag{4.2}$$

where N was the number of analysts in a group, m_q was the total number of documents for question q, and Q was the total number of questions.

The expressions for group A were:

$$F_A(jq) = \begin{cases} \frac{R_{jq}}{N} \times 100, & I_{jq} \leq R_{jq} \\ \frac{I_{jq}}{N} \times 100, & R_{jq} < I_{jq} \end{cases} \tag{4.3}$$

The values $F_A(q)$ were found by

$$F_A(q) = \frac{\sum_{j=1}^{m_q} F_A(jq)}{m_q}, \quad q = 1,2,\ldots,Q \tag{4.4}$$

Finally

$$F_A = \frac{\sum_{q=1}^{Q} F_A(q)}{Q} \tag{4.5}$$

Similar expressions were developed for group B.

The statistical test used was the Kolmogorov-Smirnov Two-Sample Test which depended on the calculation of the maximum difference, D, between the cumulative step functions of the two samples.[28] Let $S_{n1}(x)$ be the function for group A and $S_{n2}(x)$ be the function for group B.

---

[28]Siegel, op. cit., p. 128.

$S_{n1}(x)$ was defined as

$$S_{n1}(x) = \frac{K}{n_1} \tag{4.6}$$

where K was the number of scores less than or equal to a particular agreement score, x, and $n_i = Q_A$, the total number of questions analyzed by group A. Similarly

$$S_{n2}(x) = \frac{K}{n_2} \tag{4.7}$$

where K was the number of scores less than or equal to a particular agreement score, x, for $n_2 = Q_B$, the total number of questions analyzed by group B. Since group A and group B assessed the same collection of questions and documents, $n_1 = n_2 = Q$. (This condition was not necessary for this test, for there were expressions available for the condition when $n_1 \neq n_2$.[29]) Since the alternative hypothesis ($F_A \neq F_B$) did not indicate the direction of any difference in the samples, a two-tailed test was used. The expression for D was given by:

$$D = \text{maximum } |S_{n1}(x) - S_{n2}(x)| \tag{4.8}$$

Class intervals were chosen from 50 to 100 with a class length of five. If, for the value of Q, the value of D was greater than or equal to a critical value, K_D, obtained from a table of critical values,[29] then the hypothesis was rejected; otherwise, it cannot be rejected.

---

[29] Ibid., pp. 278-279.

Test 4 (Hypothesis 3)

When comparing the judgments for the two groups, there were four possible results. If the analysts of group A agreed that a document was relevant to the question, then either group B agreed that it was relevant or judged it not-relevant. These results were called R-R and R-I respectively. Similarly, group A could agree that the document was not relevant and again group B had the possible choices of R or I. These two results were labeled I-R and I-I, respectively. From the hypothesis, $p_{nc} = p_c = \frac{1}{2}$, there was an equally likely chance of change and no change in judgments for a particular document. The alternative was $p_c < p_{nc}$. A fourfold table was constructed in the following manner:

Classification of documents by group A	Classification of documents by group B: R	Classification of documents by group B: I
R	T	S
I	V	U

Figure 2. Sample Table for Judgment Differences.

where

(1) S = number of documents judged relevant by group A and not-relevant by group B.

(2) T = number of documents judged relevant by both groups.

(3) U = number of documents judged not-relevant by both groups.

(4) V = number of documents judged not-relevant by group A and relevant by group B.

From Hypothesis 3, the results that indicated a change were S and V. Using the binomial test (see Test 1), let x = S + V and N = S+T+U+V. Since N was larger than 25 and P = 1/2, the following approximation to the normal distribution was used:[30]

$$z = \frac{x - NP}{\sqrt{NPQ}} \qquad (5.1)$$

The critical value of z with $\alpha = 0.05$ was 1.65. If the z calculated by equation 5.1 was greater than or equal to this critical value, the hypothesis was rejected in favor of the alternative $p_c < p_{nc}$. If the alternative was true, any occurring change would not be significant. Hence, group A and group B could consistently agree on the relevancy of a particular document to a question.

Test 5 (Hypothesis 4)

From Hypothesis 4, the values of interest were S and V (see Figure 2). The sampling distribution was given by:

$$\chi^2 = \frac{(|S - V| - 1)^2}{S + V} \quad \text{with df} = 1.^{31} \qquad (6.1)$$

The critical values of χ^2 were obtained from a table such as that found in Siegel's text.[32] If the observed value of χ^2 obtained by Equation (6.1) was equal to or greater than the critical value of chi square shown in the table for the significance level ($\alpha = 0.05$), then Hypothesis 4 was

---

[30] Ibid., pp. 36 - 42.

[31] Ibid., p. 64.

[32] Ibid., p. 249.

rejected. This rejection would indicate that for the changes in judgments from group A to group B, there was a significant tendency for a document to be judged not relevant by the members of group A and relevant by the members of group B.

Test 6 (Hypothesis 5)

Data for testing this hypothesis were gathered by selecting at random 12 sets of 25 scores from the scores for those documents matched by random methods for group A. These were paired with another 12 samples of 25 scores likewise selected from the scores for the documents matched by maching search techniques. The Kolmogrov-Smirnov Two-Sample Test was applied to each sample in a manner similar to that described in Test 4. The alternative hypothesis was that the fraction of agreement scores was not the same for the two methods of document-question matchings.

At this point an operational hypothesis was needed to examine the results of testing Hypothesis 5 for each of the samples. Since Hypothesis 5 was either not rejected or rejected for each testing, the following hypothesis was made:

Hypothesis 5-a

If P_a is the probability of not rejecting Hypothesis 5, and P_r is the probability of rejecting it, the $P_r = P_a = 1/2$. If Hypothesis 5-a could not be rejected, then no conclusions would be made for Hypothesis 5. On the otherhand, if Hypothesis 5-a was rejected, the alternative $P_a \neq P_r$ was accepted as true. If this was the case and the number of times that Hypothesis 5 was rejected, n_r, was greater than the number of times it was not rejected, n_a, then the agreement scores for the two document selection methods of Hypothesis 5 differed significantly for

the samples tested. Similarly, if $n_r < n_a$ and the operational hypothesis was rejected, it was concluded that the agreement scores of the two retrieval methods did not exhibit any significant differences. This hypothesis (i.e., Hypothesis 5-a) was tested by the binomial test as discussed in Test 1 with $x = n_a$ if $n_a < n_r$, $x = n_r$ if $n_r \leq n_a$, $N = 12$ and $P = 1/2$.

Since the binomial test can be approximated by a normal distribution when the sample size is large and since there are expressions for testing hypotheses by the Kolmogorov-Smirnov Two-Sample Test when the sample size is larger than 25,[33] then the methods of this experimental procedure can be used to examine the properties of large samples.

---

[33] Ibid., p. 128.

CHAPTER III

SUMMARY OF DATA

Verification of Experimental Hypotheses

The judgments of relevance made by the analysts were collected, keypunched into cards, and processed by computer[34] to arrange them into response matrices[35] and to compute preliminary results. The purpose of Chapter III is to describe the important properties of the data and to discuss the testing of the hypotheses of Chapter II.

The first hypothesis considered was Hypothesis 1-a. If for a particular group (A or B) and document D_{jq}, $R_{jq} \leq 3$ or $I_{jq} \leq 3$,[36] Hypothesis 1-a was rejected for D_{jq} permitting the conclusion that $p_R \neq p_I$. The implication of this rejection was that the analysts did agree on the relevance of D_{jq} to question q. On the other hand, if Hypothesis 1-a were not rejected, then no conclusion could be made about any differences between p_R and p_I. Hence, the members of the group could not necessarily be expected to come to any agreement on the relevance of D_{jq}. As shown in Table 1, Hypothesis 1-a was not rejected 43 times for group A and 67

---

[34] Appendix B discusses the computer program and includes a copy of it.

[35] The response matrices for all questions are collected in Appendix C.

[36] This was the rejection region defined for Test 1.

times for group B.[37] The significance of these figures (i.e., the times that Hypothesis 1-a was rejected) was tested by Hypothesis 1.

The rejection region for Hypothesis 1 was established as all values of $x \leq 88$ where x was the number of times Hypothesis 1-a was not rejected for one group. From Table 1 the values of x for both groups were in the region; hence, Hypothesis 1 was rejected for both groups implying in each case that $p_o < p_1$.

Degree of agreement for the groups of analysts was obtained by evaluating Equations 4.3 to 4.5 for both groups.[38] Using the scores $F_A(q)$ and $F_B(q)$, Hypothesis 2 was tested as described in Test 3. The cumulative step functions, $S_{n1}(x)$ and $S_{n2}(x)$, and the differences, D, are given in Table 2. Since the largest difference, 3, was less than the critical difference, $K_D = 7$ for $n_1 = n_2 = 12$, Hypothesis 2 was not rejected.

The relation between the agreement scores $F_A(q)$ and $F_B(q)$ can better be seen by examining Figure 3. The scores of group A were generally higher than those of group B. Notice, that the graphs of the two groups deviate little from each other except for the documents obtained by machine searches (i.e., those documents for questions 5, 7, and 8).

By not rejecting Hypothesis 2, the indication was that there was no significant difference between the agreement scores of the two groups.

---

[37] A summary of the results for Hypothesis 1-a are found in the response matrices collected in Appendix C.

[38] These scores have been included in the response matrices of Appendix C.

Table 1. Results of Testing Hypothesis 1-a for Both Groups and All Documents

Group	Times Tested	Times Not Rejected	Times Rejected
A	202	43	159
B	202	67	135

Table 2. Cumulative Step Functions S_{n1} and S_{n2} and Differences D

	Class Intervals in Percent				
	100-96	95-91	90-86	85-81	80-76
$S_{n1}(x)$	0	3	7	11	12
$S_{n2}(x)$	0	0	5	8	9
D	0	3	2	3	3

	Class Intervals in Percent				
(Continued)	75-71	70-66	65-61	60-56	55-50
$S_{n1}(x)$	12	12	12	12	12
$S_{n2}(x)$	11	12	12	12	12
D	1	0	0	0	0

Table 3. Fourfold Table Indicating Judgment Differences Between Group A and Group B.

Judgments by Group A	Judgments by Group B R	I
R	15	0
I	31	144

However, did the two groups of analysts judge each document the same way? For example, if group A agreed that a document was relevant to a question, did group B make the same judgment? From Table 3 the groups judged a document differently only 31 times out of 190 or 16.3 per cent of the time.[39] For the 12 documents not accounted for in the table, at least one of the groups could not come to any agreement on the relevance of the document (i.e., $R_{jq} = I_{jq}$). With x = 31 and N = 190, the value of z from Equation 5.1 was 9.2. The probability of $z \geq 9.2$ was virtually zero, permitting the rejection of Hypothesis 3. Hence, for those documents on which both groups could come to some agreement on the relevance, $p_c < p_{nc}$. The indication was that when both group A and group B could come to some agreement (i.e., $R_{jq} \neq I_{jq}$ for both groups), groups A and group B could be expected to agree on the relevance of a document to a question.

If the groups differed in their judgment of a document could one difference be expected instead of another? From Table 3, there were no documents judged R by group A and I by group B, but there were 31 judged I by A and R by B. Test 5 tested the assumption of Hypothesis 4 that

[39]The decisions for each group are given in Appendix D.

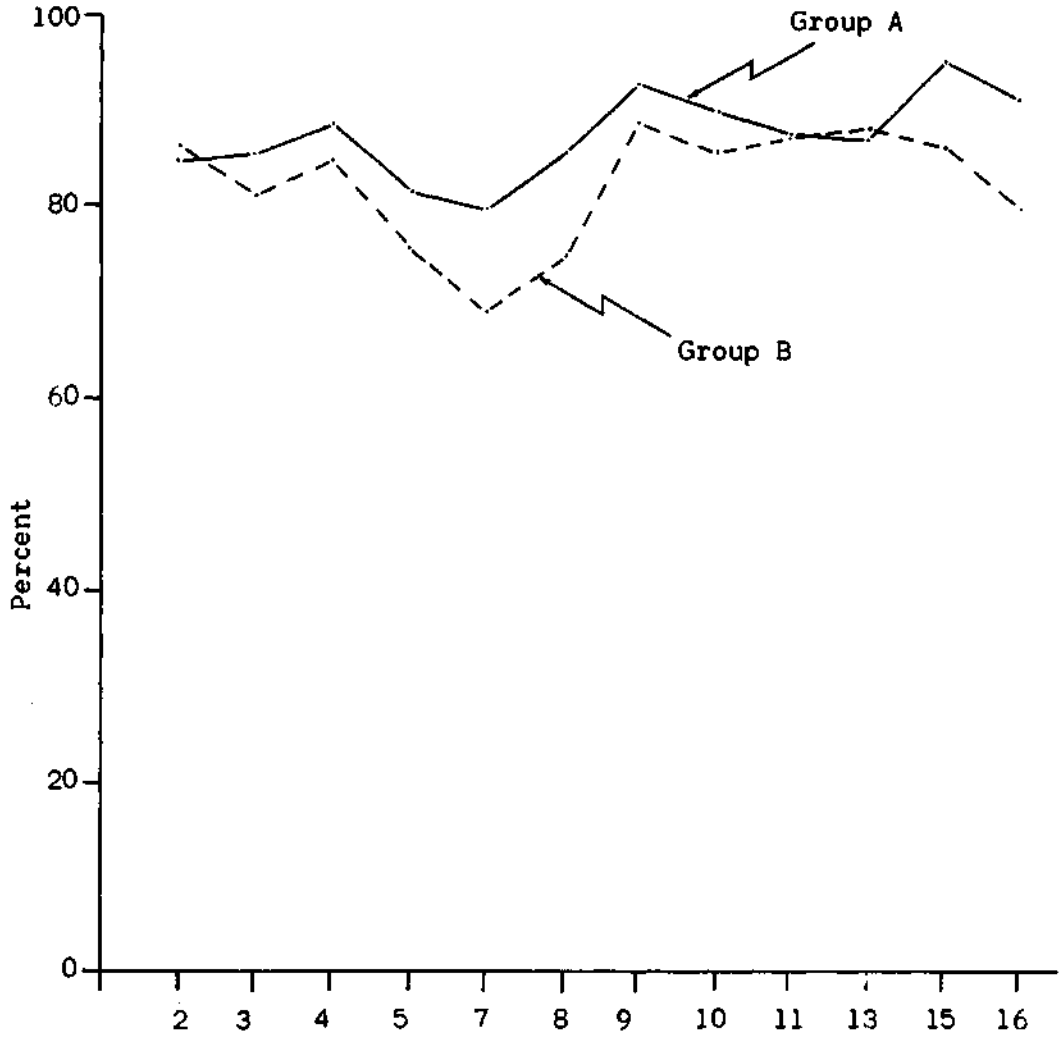


Figure 3. Average Agreement Scores Per Question.[40]

[40]There were no questions with the numbers 1, 6, 12, and 14.

$P_{RI} = P_{IR}$. Data from this experiment allowed the rejection of this hypothesis in favor of the alternative $P_{RI} < P_{IR}$. Hence, if the groups differed on their judgments of the relevancy of a document to a question, the difference expected would be an I judgment by A and an R judgment by B. But this result is somewhat questionable since one of the responses, RI, had no occurrences.

The samples described for Test 6 were selected, tabulated, and tested by the Kolmogrov-Smirnov Two-Sample Test.[41] For group A, Hypothesis 5 was rejected only twice for the 12 samples. When these results were tested by Hypothesis 5-a, the hypothesis was rejected in favor of the alternative $P_a \neq P_r$. Since $n_r < n_a$, the conclusion for Hypothesis 5 was that for group A no conclusions could be made concerning any differences between agreement scores of the machine search documents and the randomly selected documents. For group B, Hypothesis 5 was rejected for 11 of the samples, and in this case n_r was greater than n_a. Thus, the agreement scores for the two methods of document selection would be expected to differ for group B.

This concludes the discussion of the hypotheses presented in Chapter II. The remainder of the chapter reviews those aspects of the experiment which may define fruitful areas for future study.

Relevance Profile

One of the conditions of this experiment was that a document was to be judged on the basis of two mutually exclusive relevance classes --

---

[41] Appendix E contains the random samples for both groups and the cumulative step functions necessary for testing Hypothesis 5.

relevant or not-relevant. In view of the questions as to the true nature of relevance, this condition may not be valid. If so, it was expected that the experimental data might indicate the possible existence of other relevance classes; this would be determined as mentioned earlier from an analysis of the disagreement among the analysts.

Since every document reviewed by the 14 analysts of each group was evaluated as relevant (R) or not-relevant (I), there was the possibility of having from 0 to 14 R judgments for any one document. From the conditions of Hypothesis 1-a, if no more than three analysts judged a document relevant, the document was classified as not-relevant. Likewise, if at least 11 analysts within a group agreed that a document was relevant, the document was said to be a member of the relevant class. If, however, from four to ten analysts judged a document relevant, then from Hypothesis 1-a it could not be concluded that $p_R \neq p_I$; hence, these documents were not considered as members of either of the defined relevance classes. These intermediate values of R judgments were examined for the possible existence of additional relevance classes.

From the experimental data the frequency for which no members of a group judged a document relevant was found. Similar frequency counts were made for the number of R judgments from one to 14.[42] A relevance profile for the collection of documents and analysts under investigation was found by plotting a frequency distribution for the number of R judgments per document. Since the relevant and not-relevant classes were defined over several values of R judgments (0, 1, 2, and 3 for I; and 11, 12, 13, and 14 for R), the frequencies of such R judgments were considered

---

[42] These frequency counts and the computed relative frequencies have been collected into Appendix F.

collectively and plotted only once, with the not-relevant class at the left end of the distribution and the relevant class at the right. The frequencies of the R judgments of the defined classes (R and I for this experiment) were noted and the class with the minimum frequency was said to define a threshold θ. For example, if the frequency of the R class was 14 and that of the I class was 35, a θ of 14 was defined by the R class. The intermediate frequencies which exceeded θ indicate the probable existence of additional relevance classes.

The relevance profile of Group A for the entire document collection was plotted in Figure 4.[43] θ in this instance was 11 which was established by the relevant relevance class. The point designated as 1 exceeded θ, indicating the probable need for the additional relevance class of "maybe not relevant." Figure 5 presented another profile for group A, but in this case only the collection of machine searched documents was considered. The threshold of seven was larger in all cases than the intermediate frequencies. Thus, the original classes of relevant and not-relevant were sufficient in this case.

The profile of group B for all documents (Figure 6) was similar to that shown in Figure 5 which indicated that the pre-established relevance classes were probably sufficient. Group B's profile for the machine searched documents (Figure 7) was somewhat different than the others. Points 1 and 2 showed the probable need for the additional relevance classes of "maybe not relevant" and "maybe relevant" respectively.

---

[43] A profile is not included for the randomly selected documents of either group for their profiles are similar to that of all the documents of the collection.

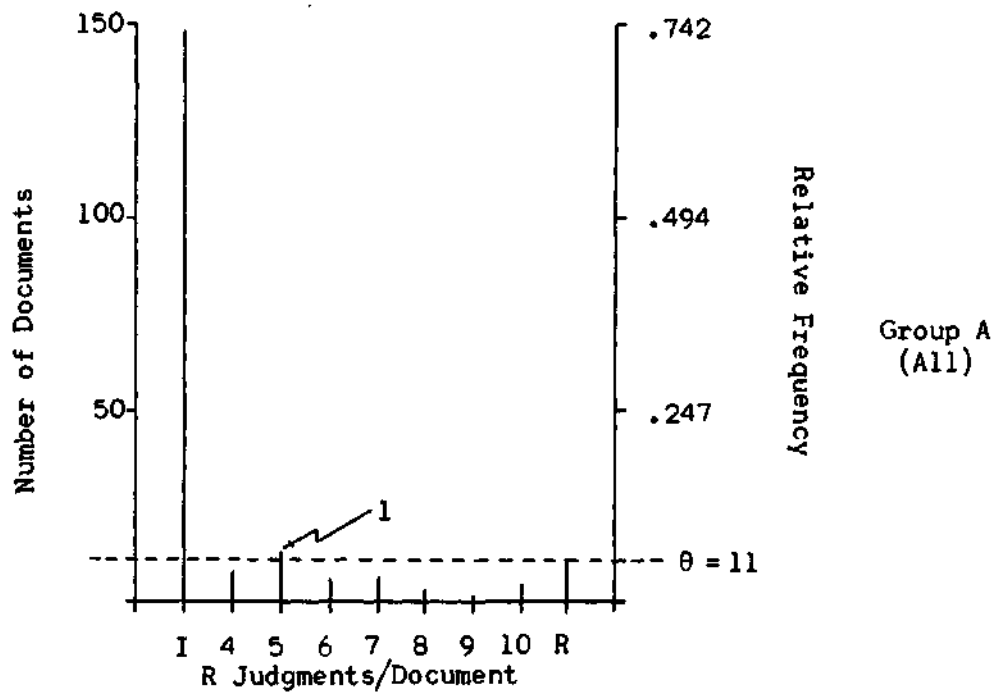


Figure 4. Relevance Profile of Group A for the Entire Document Collection.

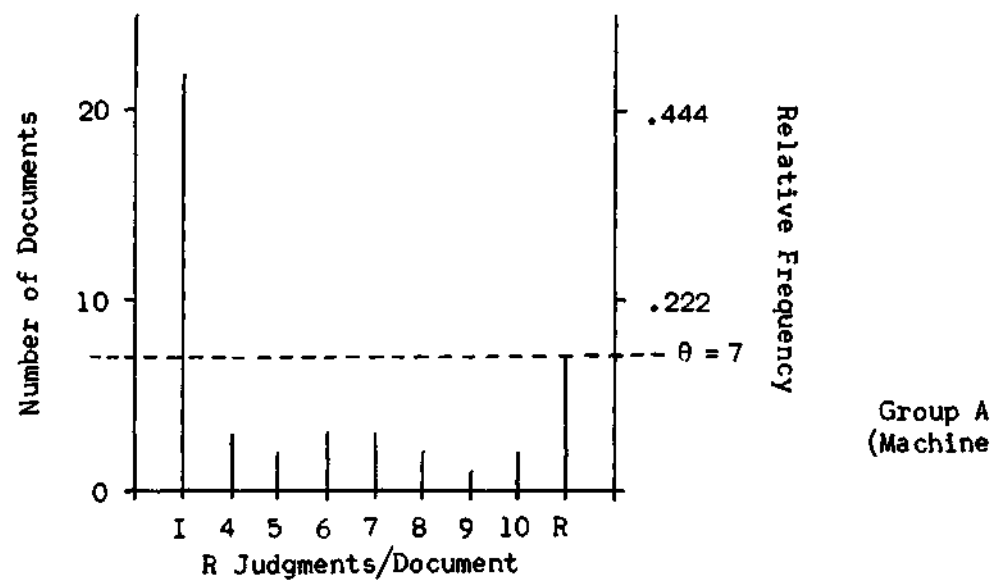


Figure 5. Relevance Profile of Group A for the Machine Retrieved Document Collection.

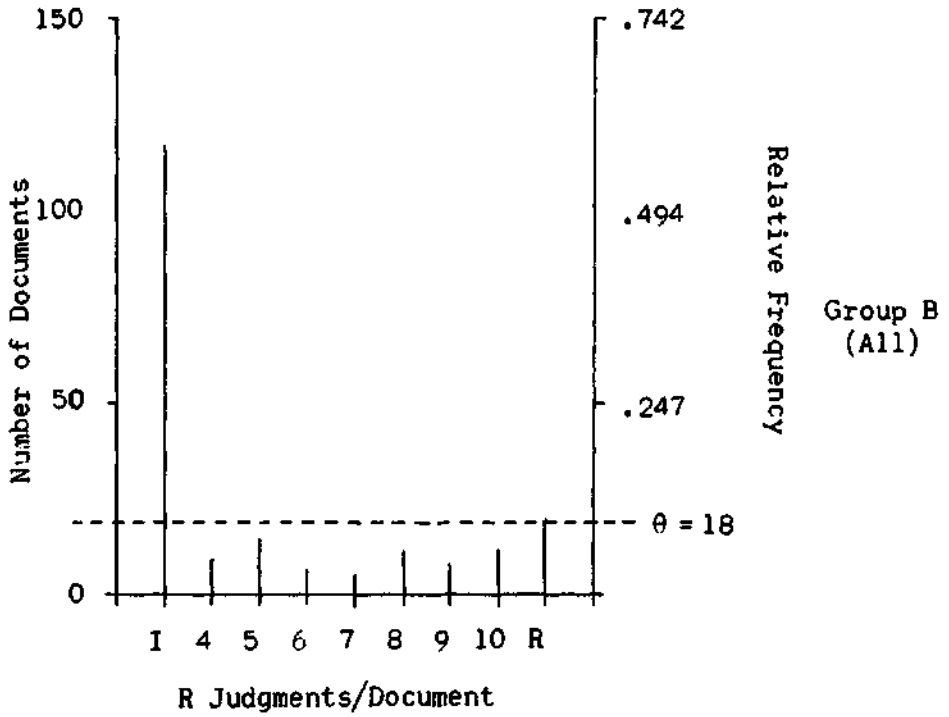


Figure 6. Relevance Profile of Group B for the Entire Document Collection

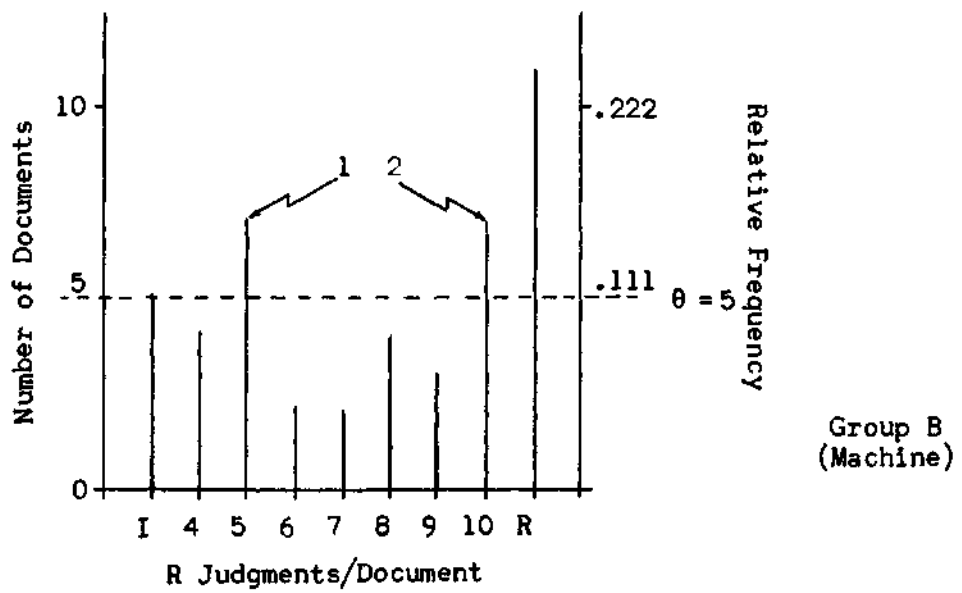


Figure 7. Relevance Profile of Group B for the Machine Retrieved Document Collection.

The relevance profile also defined a frequency distribution for the number of R judgments per document. Relative frequencies were used since they were in fact empirical probabilities used to estimate the probability of having a given number of analysts judge any document of the collection relevant to a given set of questions. If these probabilities were known for the collection of documents of an information system, then for a given set of questions, it would be possible to predict what fraction of the system users would agree that a document was relevant. For the analysts, questions, and documents of this test, the frequency distributions were defined by Figure 4, Figure 5, Figure 6, and Figure 7 with the relative frequencies shown on the right ordinate scale.

Assessment Times

As the analysts reviewed and judged the relevance of the documents, they were asked to record for each question the time necessary to make the required judgments.[44] The times were recorded for each question and an average time was obtained for each question. By dividing the average time per question by the number of documents per question, the average time per document for that question was obtained. From Figure 8 notice that the members of group A could always make their judgments quicker than the members of group B. This appeared understandable since the non-experts of group B were not expected to be as familiar with the subject field as the experts of group A and thus were likely to require more time to acquire an understanding of the document contents before making a judgment.

---

[44]The times that the analysts took to make their judgments are given in the tables of Appendix G.

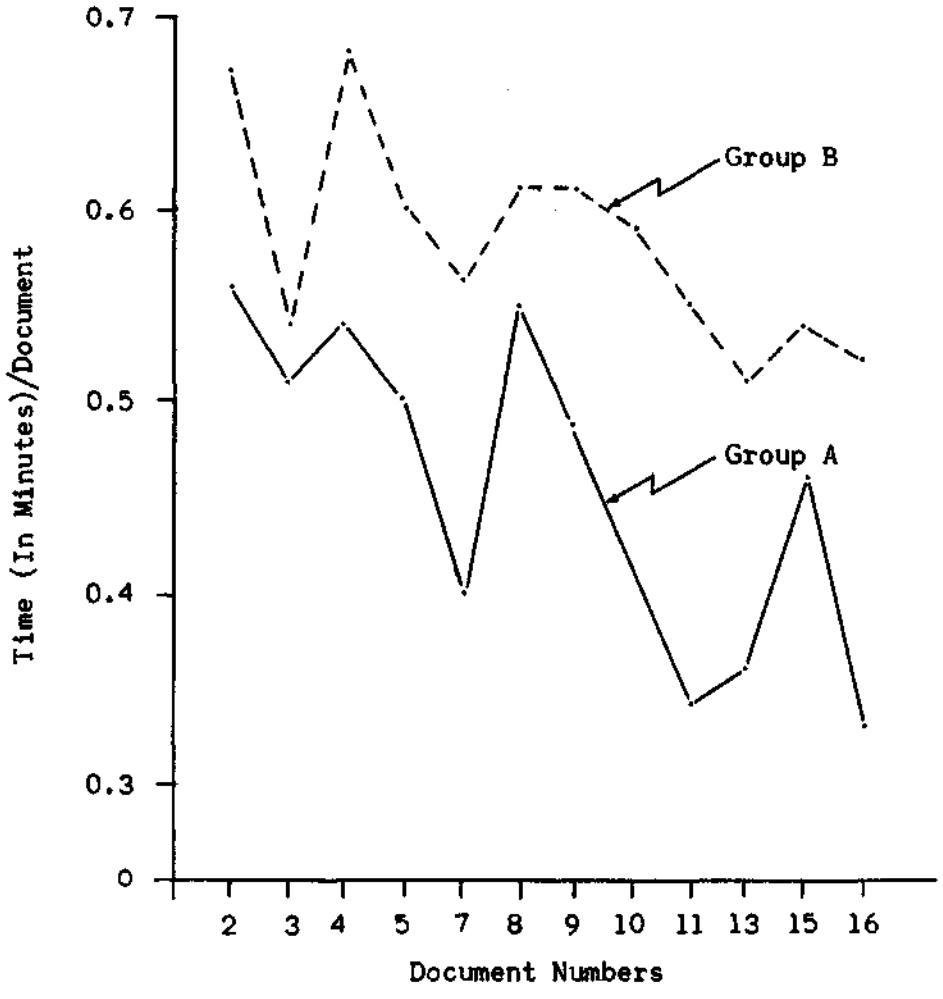


Figure 8. Judgment Time Per Document for Each Question and Group.

An interesting result can be seen by examining Figure 9 where the average times per document were plotted against the number of documents per question. For both groups as the number of documents per question increased, the time per document also increased until a maximum time was reached at 15 documents per question. In the interval from 16 to 25 documents per question, the average time per document was generally decreasing. It appeared, then, that if no more than 15 document abstracts per question were submitted to a group of analysts for examination and assessment of relevance, the analysts would tend to spend more time per document abstract as the number of abstracts per question increased. However, if more than 15 abstracts were included with each question, the analysts would spend less time per abstract.

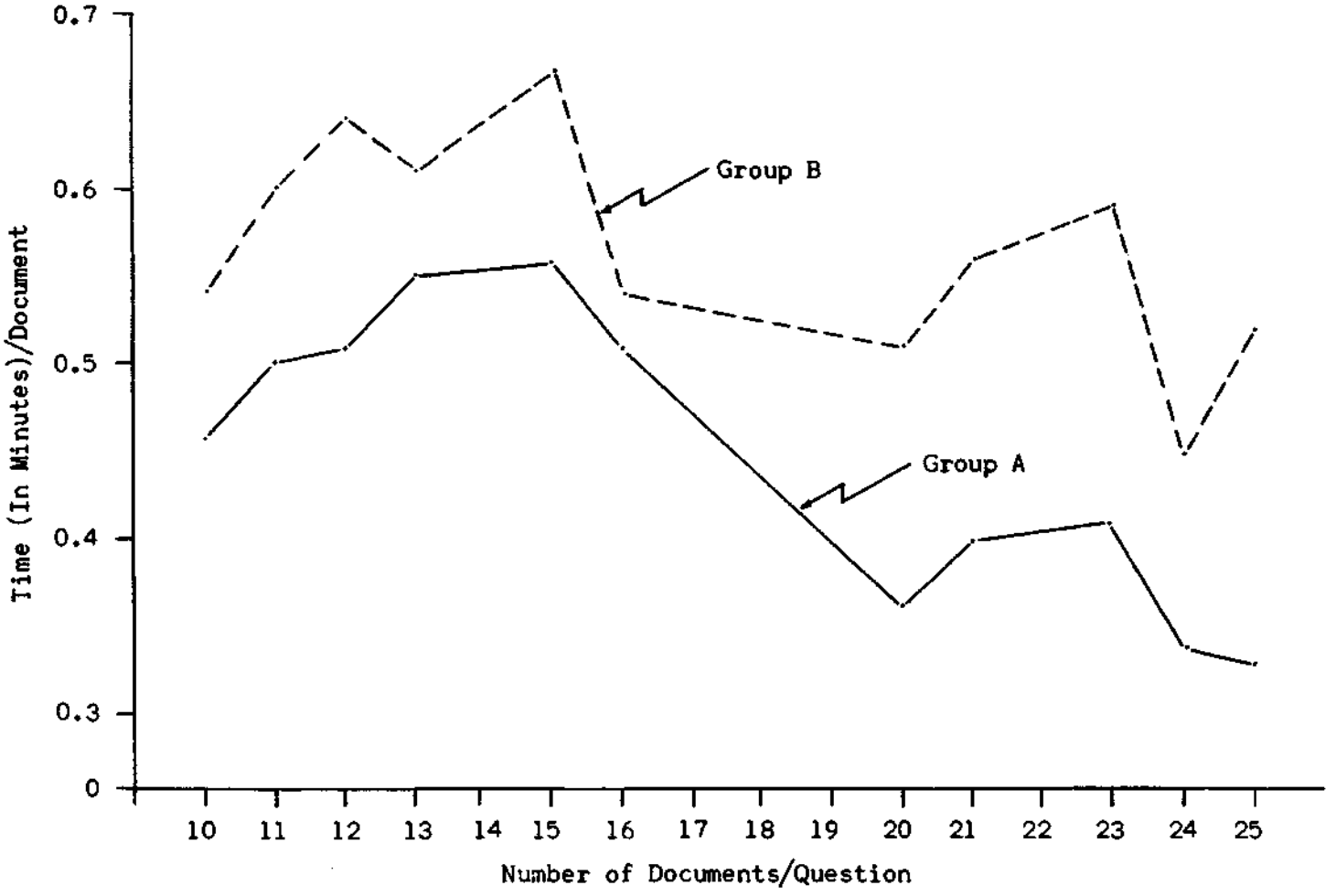


Figure 9. Judgment Time Per Document Vs Number of Documents Per Question.

CHAPTER IV

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

The purpose of this experiment was twofold: (1) To design an experimental procedure for testing inter- and intra-group consistency of judgments of the relevancy of documents to specific questions by subject experts and non-experts, and to develop valid statistical procedures for analyzing the data collected; and (2) to perform a pilot experiment testing these procedures.

An experimental procedure with the following constraints was designed to test the inter- and intra-group consistency of relevance judgments:

(1) only two groups of document analysts (experts and non-experts) were used;

(2) relevance was understood such that a document was judged either relevant or not-relevant to a question;

(3) the search questions were formulated by the members of the expert group;

(4) the documents were matched with the search questions both by random selection from an abstract journal (STAR) and by machine searches by the Defense Documentation Center.

The statistical procedures for testing the hypotheses were of general utility under the experimental constraints and produced consistent

results for both classes of experimental data.

The conclusions and implications of the experimental hypotheses and data were:

(1) From Hypothesis 1, the members of both groups could consistently agree on the relevance of documents to questions.

(2) The consistency of judgments by group B did not differ significantly from that of group A.

(3) The method of document selection (i.e., random and machine matching of documents to questions) did not generally affect the degree of agreement of group A, but for group B the results of testing Hypothesis 5 indicated that the agreement scores for the two selection methods differed significantly. From Figure 3, the scores for the random searches were in every case lower than those for the machine matched documents.

(4) If the groups differed on their judgments of the relevance of a document, testing of Hypothesis 4 indicated that the difference expected would be an I judgment by group A and an R judgment by group B for D_{jq}.

(5) Since $F_B < F_A$, group A was more sure of their judgments of relevance than group B.

(6) From Figure 9, the time required for the analysts of group A to make their assessments of relevance was, on the average, less than that of the analysts of group B.

(7) As shown by the relevance profiles, the preestablished relevance classes (relevant and not-relevant) were sufficient in two cases, one for group A and the other for group B; but in the other two cases

considered, group B required two intermediate relevance classes where group A required at most one.

(8) Also from the relevance profiles, the optimum conditions for assessing the relevance of documents to questions would be to require subject experts to assess machine retrieved documents on the basis of the two mutually exclusive relevance classes -- relevant and not-relevant.

Recommendations

This experiment was just a beginning in investigating the nature of human relevance assessments of documents. The experimental procedure was tested only by a pilot experiment from which some preliminary conclusions concerning the characteristics of relevance assessments and the properties of relevance were reached. Due to the nature of statistics, these conclusions were restricted to the conditions of the pilot test. Since the data analysis uncovered several important aspects which were worthy of further study, the following suggestions seem worthwhile:[45]

(1) The pilot experiment should be repeated with the same analysts, questions, and documents with the new relevance classes as defined by the relevance profiles.

(2) The experiment should be repeated on a wider scope to include:

a. larger groups of analysts

b. analysts with various levels of education and training in the subject fields

---

[45]These suggestions do not extend beyond the boundaries of this experiment, since many other areas of future study and experimentation have been previously described in other literature.

c. analysts with different degrees of interest and relationship to the subject field and to the search questions

d. a wide variety of subject fields

e. different forms of document surrogates and total text

(3) The "relative frequency" hypothesis presented in Chapter III must be investigated to determine its validity.

(4) In future experiments valuable information could be obtained from post-analysis conferences with the analysts to determine the properties of the documents, questions, and subject field that prompted the analysts to make their respective assessments.

(5) To make future experiments more realistic, the documents submitted for assessments should be retrieved by mechanical search methods.

APPENDIX A

The following is a summary of the important notations used throughout the text.

A - a group of subject experts

B - a group of non-experts in a given subject area

q - a question

i - an analyst

j - a document

D_{jq} - document j submitted with question q

N - the number of analysts

Q_A - the total number of questions analyzed by group A

Q_B - the total number of questions analyzed by group B

m_q - the total number of documents submitted with question q

θ_{ijq} - the response (R or I) of analyst i to document j, question q

φ_{ijq} - 1 when θ_{ijq} is R and 0 when θ_{ijq} is I

R - response assigned to D_{jq} when judged by an analyst relevant to question q

I - response assigned to D_{jq} when judged by an analyst relevant to question q

R_{jq} - total number of relevant responses (R) for a group (A or B) and D_{jq}

I_{jq} - total number of not-relevant responses (I) for a group (A or B) and D_{jq}

$F_A(jq)$ - the fraction of the members of group A who agreed on the relevancy of D_{jq} to q

$F_B(jq)$ - same as $F_A(jq)$ except for group B

$F_A(q)$ - the fraction of the members of group A who agreed on the relevance judgments for q

$F_B(q)$ - same as $F_A(q)$ except for group B

F_A - the fraction of the members of group A who agreed on relevance judgments for Q_A

F_B - same as F_A except for group B

p_{nc} - the probability that group A and group B made the same relevance judgments for a document

p_c - the probability that group A and group B differed on relevance judgments for a document

P_{RI} - the probability that a document was judged R by A and I by B

P_{IR} - the probability that a document was judged I by A and R by B

p_o - the probability of not rejecting Hypothesis 1-a

p_1 - the probability of rejecting Hypothesis 1-a

n_a - the number of times that Hypothesis 5-a was not rejected

n_r - the number of times that Hypothesis 5-a was rejected

P_a - the probability of not rejecting Hypothesis 5

P_r - the probability of rejecting Hypothesis 5

APPENDIX B

The computer program written in ALGOL 60 for implementation on the Burroughs B-5500 computer provided instructions for sorting the raw data and for arranging the data and preliminary results into response matrices. The functions of the program have been explained by means of a brief flow chart and comment statements designated by a "%" which are located within the program.

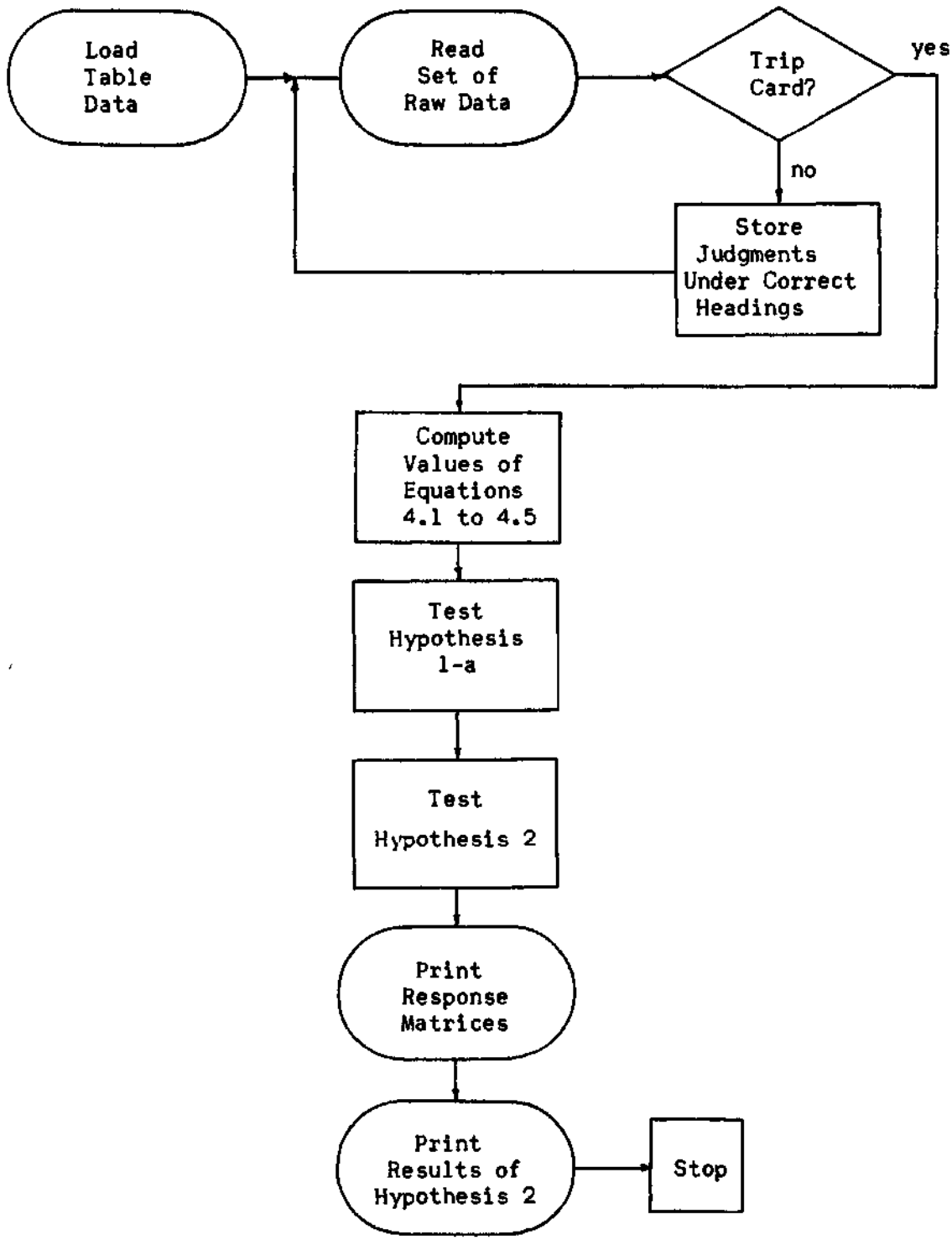


Figure 10. Computer Program Flow Diagram.

```
          BEGIN
FILE IN          CARD (2,10);
FILE OUT         LINE 1(2,15);
INTEGER          B,C,J,K,I,N,XA,T,LD,DA,DB,ANA,Q,G;
REAL             TA,TB,AR,AI,BR,BI;
ALPHA            GR;
INTEGER ARRAY    TTQ,TTA,TQ[0:2],TOD,QUE[0:16,0:2],INT,CA,CB,DAB[0:10],
                 NT,D[0:30],AN[0:20,0:2],REJA,REJB[0:16,0:25],AT,BT[0:16,
                 0:25,0:31],NA,VV[0:25,0:2];
REAL ARRAY       PA,PB[0:16],AA,BB[0:16,0:25];
ALPHA ARRAY      JUD[0:25],DOC[0:16,0:25,0:2],MA,MB[0:20,0:25,0:16];
LABEL            R,L1,L2,L3,L4,L5,L6,L7,L8,L9,ST,L10,L11,L12,L13,L14,L15,
                 L16,L17,L26,L26A,L26B,L27,L27A,L28,L28A,L29,L30,L31,L32,
                 L33,PR,L40,L41,L42,L43;
% THE FOLLOWING LIST AND FORMAT STATEMENTS ARE FOR PROGRAM INPUT. THE
% FORMAT FOR INPUT CARDS CAN BE FOUND BY EXAMINING THE READ STATEMENTS
% USING THESE LISTS AND FORMATS. AN EXPLANATION OF THE VARIABLES USED
% IS AS FOLLOWS:
%                TTQ[1] AND TTQ[2] = TOTAL QUESTIONS FOR GROUP A AND GROUP
%                          B RESPECTIVELY
%                TOD[K,J] = TOTAL DOCUMENTS FOR GROUP J (J=1,2) QUESTION K
%                DOC[K,N,J] = DOCUMENT NUMBERS
%                INT[I] = THE LOWER CLASS BOUNDARIES FOR TESTING
%                          HYPOTHESIS 2
%                NT[I] AND D[I] = THE SAMPLE SIZE AND CRITICAL VALUES
%                          RESPECTIVELY FOR TESTING HYPOTHESIS 2
%                TTA[1] AND TTA[2] = TOTAL ANALYSTS FOR GROUP A AND GROUP
%                          B RESPECTIVELY
%                AN[I,J] = ANALYST NUMBERS FOR GROUP J (J=1,2)
%                QUE[I,J] = QUESTION NUMBERS
%                GR, ANA, Q = THIS IS FOR INPUT OF EXPERIMENTAL DATA WHERE
%                          GR IS THE GROUP (A OR B), ANA THE ANALYST NUMBER
%                          , AND Q THE QUESTION NUMBER
%                JUD[N] = THE R OR I JUDGMENTS FOR A PARTICULAR GR, ANA,
%                          AND Q
```

```
%              NA[I,T] AND VV[I,T] = THE SAMPLE SIZE AND CRITICAL VALUES
%                       FOR BINOMIAL TEST. T=1 FOR TWO TAIL; T=2 FOR ONE
%                       TAIL TEST.
LIST           LT29(TTQ[1],TTQ[2]);
LIST           LT30(FOR J←1 STEP 1 UNTIL 2 DO FOR K←1 STEP 1 UNTIL
               TTQ[J] DO TOD[K,J]);
LIST           LT31(FOR J←1 STEP 1 UNTIL 2 DO FOR K←1 STEP 1 UNTIL
               TTQ[J] DO FOR N←1 STEP 1 UNTIL TOD[K,J] DO DOC[K,N,J]);
LIST           LT32(FOR I←1 STEP 1 UNTIL 10 DO INT[I]);
LIST           LT33(NT[I],O[I]);
LIST           LT50(TTA[1],TTA[2]);
LIST           LT51(FOR J←1 STEP 1 UNTIL 2 DO FOR I←1 STEP 1 UNTIL
               TTA[J] DO AN[I,J]);
LIST           LT52(FOR J←1 STEP 1 UNTIL 2 DO FOR I←1 STEP 1 UNTIL
               TTQ[J] DO QUE[I,J]);
LIST           LT1(GR,ANA,Q);
LIST           LT34(FOR N←1 STEP 1 UNTIL TOD[K,J] DO JUD[N]);
LIST           LT60(NA[I,T],VV[I,T]);
FORMAT         FM2("ERROR: TQ[1] ≠ TQ[2]");
FORMAT         FM29(2I3);
FORMAT         FM30(20I3);
FORMAT         FM31(20A4);
FORMAT         FM32(10I3);
FORMAT         FM33(I3,I2);
FORMAT         FM1(A1,X3,I3,X2,I3);
FORMAT         FM34(30(A1,X1));
FORMAT         FM50(2I3);
FORMAT         FM51(20(X1,I3));
FORMAT         FM52(20I3);
FORMAT         FM60(2I2);
% THE NEXT SECTION OF PROGRAM CONSISTS OF FORMATS AND PROCEDURES TO
% PRINT THE RESULTS OF COMPUTATIONS PERFORMED BY THIS PROGRAM. THE
% PROCEDURES HAVE THE FOLLOWING FUNCTIONS:
%              HEAD AND HEADC - PRINTS THE HEADING AND CONTINUED
%                       HEADING GIVING GROUP AND QUESTION NUMBER
```

```
%              RR(B,C) - PRINTS THE R AND I RESPONSES
%              SUM(B,C) - PRINTS R(J,Q) AND I(J,Q)
%              AGS(B,C) - PRINTS THE AGREEMENT SCORES F, F(Q),AND
%                         F (J,Q)
%              H13(B,C) - PRINTS THE RESULTS OF TESTING HYPOTHESIS 1-A
%              HYP2 - PRINTS THE RESULTS OF HYPOTHESIS 2
PROCEDURE      HEAD;
          BEGIN
LIST           LT2(QUE[K,J]);
FORMAT         FM3A("GROUP  A"/"QUESTION",X2,I2);
FORMAT         FM3B("GROUP  B"/"QUESTION",X2,I2);
FORMAT         FM4(/X30,"DOCUMENT NUMBERS"/);
               IF J=1 THEN WRITE(LINE,FM3A,LT2) ELSE WRITE(LINE,FM3B,
               LT2); WRITE(LINE,FM4);
          END OF HEAD;
PROCEDURE      HEADC;
          BEGIN
LIST           LT2(QUE[K,J]);
FORMAT         FM3A("GROUP  A"/"QUESTION",X2,I2," - CONTINUED");
FORMAT         FM3B("GROUP  B"/"QUESTION",X2,I2," - CONTINUED");
FORMAT         FM4(/X30,"DOCUMENT NUMBERS"/);
               IF J=1 THEN WRITE(LINE,FM3A,LT2) ELSE WRITE(LINE,FM3B,
               LT2); WRITE(LINE,FM4);
          END OF HEADC;
PROCEDURE      RR(B,C);  INTEGER B,C;
          BEGIN
INTEGER        I,N;
LIST           LT3(FOR N←C STEP 1 UNTIL B DO DOC[K,N,J]);
LIST           LT4A(AN[I,J],FOR N←C STEP 1 UNTIL B DO MA[I,N,K]);
LIST           LT4B(AN[I,J],FOR N←C STEP 1 UNTIL B DO MB[I,N,K]);
FORMAT         FM5("ANALYST",X2,13(A4,X1));
FORMAT         FM6(X2,I3,X6,13(A1,X4));
               WRITE(LINE,FM5,LT3); WRITE(LINE,DASH1); FOR I←1 STEP 1
               UNTIL TTA[J] DO IF J=1 THEN WRITE(LINE,FM6,LT4A) ELSE
               WRITE(LINE,FM6,LT4B);
```

```
          END OF RR(B,C);
PROCEDURE      SUM(B,C); INTEGER B,C;
          BEGIN
LIST           LT2(FOR N←C STEP 1 UNTIL B DO BT[K,N,I]);
LIST           LT1(FOR N←C STEP 1 UNTIL B DO AT[K,N,I]);
FORMAT         FM3("I(J,Q)",X4,13(I2,X3));
FORMAT         FM2("R(J,Q)",X4,13(I2,X3));
               WRITE(LINE,DASH1); IF J=1 THEN BEGIN I←2; WRITE(LINE,
               FM2,LT1); I←3; WRITE(LINE,FM3,LT1); WRITE(LINE,DASH1);
               END ELSE BEGIN I←2; WRITE(LINE,FM2,LT2); I←3; WRITE(
               LINE,FM3,LT2);WRITE(LINE,DASH1); END;
          END OF SUM(B,C);
PROCEDURE      AGS(B,C); INTEGER B,C;
          BEGIN
LIST           LT4A(FOR N←C STEP 1 UNTIL B DO AA[K,N]);
LIST           LT4B(FOR N←C STEP 1 UNTIL B DO BB[K,N]);
LIST           LT5A(PA[K]);
LIST           LT5B(PB[K]);
LIST           LT6A(TA);
LIST           LT6B(TB);
FORMAT         FM4("F (J,Q)",X2,13(F4.1,X1));
FORMAT         FM6("F (Q)",X10,F4.1);
FORMAT         FM8("F ",X13,F4.1);
               IF J=1 THEN WRITE(LINE,FM4,LT4A) ELSE WRITE(LINE,FM4,
               LT4B); WRITE(LINE,DASH1); IF B=TOD[K,J] THEN BEGIN IF
               J=1 THEN WRITE(LINE,FM6,LT5A) ELSE WRITE(LINE,FM6,LT5B);
               WRITE(LINE,DASH1); IF K=TQ[J] THEN BEGIN IF J=1 THEN
               WRITE(LINE,FM8,LT6A) ELSE WRITE(LINE,FM8,LT6B); WRITE(
               LINE,DASH1); END; END;
          END OF AGS(B,C);
PROCEDURE      H13(B,C); INTEGER B,C;
          BEGIN
LIST           LT4A(FOR N←C STEP 1 UNTIL B DO REJA[K,N]);
LIST           LT4B(FOR N←C STEP 1 UNTIL B DO REJB[K,N]);
FORMAT         FM4("HYP 1-A",X4,13(I1,X4));
```

```
                IF J=1 THEN WRITE(LINE,FM4,LT4A) ELSE WRITE(LINE,FM4,
                LT4B); WRITE(LINE,DASH1); WRITE(LINE,DASH6);WRITE(LINE
                [PAGE]);
            END OF H13(B,C);
PROCEDURE       HYP2;
            BEGIN
INTEGER         I ;
LIST            LT1(FOR I=1 STEP 1 UNTIL 10 DO CA[I]);
LIST            LT4(LD)   ;
LIST            LT2(FOR I=1 STEP 1 UNTIL 10 DO CB[I]);
LIST            LT3(FOR I=1 STEP 1 UNTIL 10 DO DAB[I]);
FORMAT          FM2(X10,"GROUP B",X6,10(I1,X5)/);
FORMAT          FM1(X10,"GROUP A",X6,10(I1,X5));
FORMAT          FM3(X10,"A - B",X8,10(I1,X5));
FORMAT          FM8(//X10,"THE LARGEST DIFFERENCE IS",I2) ;
FORMAT          FM4(X29,"HYPOTHESIS 2 IS NOT REJECTED")        ;
FORMAT          FM5(X29,"HYPOTHESIS 2 IS REJECTED");
FORMAT          FM6(X44,"HYPOTHESIS 2"/);
FORMAT      FM7(X10,"PERCENT   100-96 95-91 90-86 85-81 80-76 75-71 70-66
 65-61 60-56 55-50"/)  ;
                WRITE(LINE,FM6);
                WRITE(LINE,FM7)   ;
                WRITE(LINE,FM1,LT1);
                WRITE(LINE,FM2,LT2);
                WRITE(LINE,FM3,LT3);
                WRITE(LINE[DBL]);
                WRITE(LINE,FM8,LT4)   ;
                IF G=0 THEN WRITE(LINE,FM5)
                ELSE WRITE(LINE,FM4);
            END OF HYP2;
% THE PROCEDURE REJECT HAS THE VALUE OF 0 IF HYPOTHESIS 1-A IS NOT
% REJECTED AND 1 IF IT IS REJECTED.
INTEGER PROCEDURE REJECT(X,NN,T);  INTEGER X,NN,T;
            BEGIN
INTEGER         I;
```

```
LABEL           L1,FIN;
FORMAT          FM1("REJECT ERROR");
          BEGIN I←1;
L1:             IF NA[I,T]=NN THEN BEGIN IF X>VV[I,T] THEN REJECT←0 ELSE
                REJECT←1; GO TO FIN; END; IF I=21 THEN BEGIN WRITE(LINE,
                FM1); GO TO FIN; END;  I←I+1; GO TO L1; END;
FIN:            END OF REJECT;
% THIS BEGINS THE ACTUAL PROGRAM. THE FOLLOWING READ STATEMENTS ARE
% EXPLAINED IN THE LIST STATEMENT REFERENCED BY THE RESPECTIVE READ.
                READ(CARD,FM29,LT29);READ(CARD,FM30,LT30);READ(CARD,FM31,
                LT31);READ(CARD,FM32,LT32);FOR I←1 STEP 1 UNTIL 28 DO
                READ(CARD,FM33,LT33); READ(CARD,FM50,LT50); READ(CARD,
                FM51,LT51); READ(CARD,FM52,LT52); TQ[1]←TTQ[1];
                TQ[2]←TTQ[2];
R:              READ(CARD,FM1,LT1); IF GR="A" THEN BEGIN J←1;GO TO L1;
                END; IF GR="C" THEN GO TO ST; J←2;
% A TRIP CARD WITH A "C" FOLLOWS ALL EXPERIMENTAL DATA.
L1:             I←1;
L2:             IF AN[I,J]=ANA THEN GO TO L3; I←I+1; GO TO L2;
L3:             K←1;
L5:             IF QUE[K,J]=Q THEN GO TO L4; K←K+1; GO TO L5;
% STORE JUDGMENTS IN MA OR MB AND COUNT FREQUENCY OF R AND I
L4:             READ(CARD,FM34,LT34); IF J=1 THEN BEGIN N←1;
L7:             MA[I,N,K]←JUD[N]; AT[K,N,1]←AT[K,N,1]+1; IF JUD[N]="R"
                THEN AT[K,N,2]←AT[K,N,2]+1 ELSE AT[K,N,3]←AT[K,N,3]+1;
                IF N=TOD[K,J] THEN GO TO R; N←N+1; GO TO L7; END ELSE
                BEGIN N←1;
L6:             MB[I,N,K]←JUD[N]; BT[K,N,1]←BT[K,N,1]+1; IF JUD[N]="R"
                THEN BT[K,N,2]←BT[K,N,2]+1 ELSE BT[K,N,3]←BT[K,N,3]+1;
                IF N=TOD[K,J] THEN GO TO R; N←N+1; GO TO L6; END;
% COMPUTE AGREEMENT SCORES FA, FA(Q), AND FA(J,Q).
ST:             TA←0.0; J←1; K←1;
L8:             N←1;
L9:             AR←AT[K,N,2]/AT[K,N,1]; AI←AT[K,N,3]/AT[K,N,1]; IF AR≥AI
                THEN BEGIN AA[K,N]←AR×100.0; PA[K]←PA[K]+AA[K,N]; GO TO
```

```
                L10; END ELSE BEGIN AA[K,N]←AI×100.0; PA[K]←PA[K]+AA[K,N]
                ; GO TO L10; END;
L10:            IF N=TOD[K,J] THEN GO TO L11; N←N+1; GO TO L9;
L11:            PA[K]←PA[K]/TOD[K,J]; TA←TA+PA[K]; IF K=TQ[J] THEN GO TO
                L12; K←K+1; GO TO L8;
L12:            TA←TA/TQ[J];
% COMPUTE AGREEMENT SCORES FB, FB(Q), AND FB(J,Q).
                TB←0.0; J←2; K←1;
L13:            N←1;
L14:            BR←BT[K,N,2]/BT[K,N,1]; BI←BT[K,N,3]/BT[K,N,1]; IF BR≥BI
                THEN BEGIN BB[K,N]←BR×100.0; PB[K]←PB[K]+BB[K,N]; GO TO
                L15; END ELSE BEGIN BB[K,N]←BI×100.0; PB[K]←PB[K]+BB[K,N]
                ; GO TO L15; END;
L15:            IF N=TOD[K,J] THEN GO TO L16; N←N+1; GO TO L14;
L16:            PB[K]←PB[K]/TOD[K,J]; TB←TB+PB[K]; IF K=TQ[J] THEN GO TO
                L17; K←K+1; GO TO L13;
L17:            TB←TB/TQ[J];
% TESTING OF HYPOTHESIS 1=A FOR BOTH GROUPS. THE READ LOADS THE SAMPLE
% SIZE AND CRITICAL VALUES FOR THE BINOMIAL TEST. THE CLOSE STATEMENT
% PERMITS MULTIPROCESSING.
                FOR T←1 STEP 1 UNTIL 2 DO FOR I←1 STEP 1 UNTIL 21 DO
                READ(CARD,FM60,LT60); CLOSE(CARD,RELEASE); T←1; FOR J←1
                STEP 1 UNTIL 2 DO FOR K←1 STEP 1 UNTIL TQ[J] DO FOR N←1
                STEP 1 UNTIL TOD[K,J] DO IF J=1 THEN BEGIN IF AT[K,N,2]≥
                AT[K,N,3] THEN XA←AT[K,N,3] ELSE XA←AT[K,N,2]; REJA[K,N]←
                REJECT(XA,AT[K,N,1],T); END ELSE BEGIN IF BT[K,N,2]≥BT[K,
                N,3] THEN XA←BT[K,N,3] ELSE XA←BT[K,N,2]; REJB[K,N]←
                REJECT(XA,BT[K,N,1],T); END;
% TEST HYPOTHESIS 2.
                IF TQ[1] ≠ TQ[2] THEN BEGIN WRITE(LINE,FM2); GO TO PR;
                END; T←TQ[1]; K←1;
L26:            I←1;
L27:            IF PA[K]≥INT[I] THEN BEGIN CA[I]←CA[I]+1; GO TO L28; END;
                I←I+1; GO TO L27;
L28:            IF K=T THEN GO TO L26A; K←K+1; GO TO L26;
```

```
L26A:           K←1;
L26B:           I←1;
L27A:           IF PB[K]≥INT[I] THEN BEGIN CB[I]←CB[I]+1; GO TO L28A;
                END; I←I+1; GO TO L27A;
L28A:           IF K=T THEN GO TO L29; K←K+1; GO TO L26B;
L29:            LD←0; DA←0; DB←0; I←1;
L30:            DA←DA+CA[I]; CA[I]←DA; DB←DB+CB[I]; CB[I]←DB; DAB[I]←
                CA[I]-CB[I]; IF DAB[I]<0 THEN DAB[I]←(-1)×DAB[I]; IF LD<
                DAB[I] THEN LD←DAB[I]; IF I=10 THEN GO TO L31; I←I+1; GO
                TO L30;
L31:            I←1;
L33:            IF T=NT[I] THEN GO TO L32; I←I+1; GO TO L33;
L32:            IF LD<D[I] THEN G←1 ELSE G←0;
PR:             J←1;
L42:            K←1;
L41:            IF TOD[K,J]<13 THEN BEGIN C←1; B←13; HEAD; RR(B,C); SUM(B
                ,C); AGS(B,C); H13(B,C); C←14; B←TOD[K,J]; HEADC; RR(B,C)
                ; SUM(B,C); AGS(B,C); H13(B,C); END ELSE BEGIN C←1; B←
                TOD[K,J]; HEAD; RR(B,C); SUM(B,C); AGS(B,C); H13(B,C);
                END; IF K=TQ[J] THEN GO TO L40; K←K+1; GO TO L41;
L40:            IF J=2 THEN GO TO L43; J←2; GO TO L42;
L43:            HYP2;
        END.
```

APPENDIX C

This appendix contains the response matrices for each question for both groups of analysts. Each matrix shows the R and I responses by analyst to each document; the R_{jq} and I_{jq} total; and the values of Equations 4.1 to 4.5 where applicable. When one of these equations evaluated to 100.0, the symbol "****" appeared in the matrix. A "1" recorded for "HYP 1-A" implied that Hypothesis 1-a was rejected for a document and a "0" implied that the hypothesis was not rejected.

GROUP A
QUESTION 2

DOCUMENT NUMBERS

ANALYST	B420	B910	B200	B460	B220	B580	B600	A580	A410	A550	B770	A570	A480
101	I	R	I	R	I	I	I	R	I	R	R	I	I
102	I	I	I	I	I	I	I	I	I	I	I	I	I
103	I	I	I	I	I	I	R	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	I	R	I	I
105	R	I	I	I	I	I	R	R	I	I	R	I	I
106	R	R	I	I	I	I	R	I	I	I	I	I	I
107	I	R	I	I	I	I	R	I	I	I	I	I	I
108	I	I	I	I	I	I	I	I	I	I	I	I	I
109	I	R	I	I	I	I	I	I	I	I	I	I	I
110	I	I	I	I	I	I	I	I	I	I	R	I	I
111	I	R	I	I	I	I	R	I	I	I	I	I	I
112	I	R	I	I	I	I	I	I	I	I	I	I	I
113	I	R	I	I	R	I	I	I	I	I	R	I	I
114	I	I	I	I	R	I	R	I	I	I	R	I	I
R(J,Q)	2	7	0	1	2	0	6	2	0	1	6	0	0
I(J,Q)	12	7	14	13	12	14	8	12	14	13	8	14	14
F (J,Q)	85.7	50.0	****	92.9	85.7	****	57.1	85.7	****	92.9	57.1	****	****
HYP 1=A	1	0	1	1	1	1	0	1	1	1	0	1	1
			*				*				*		

GROUP A
QUESTION 2 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A210	A970
101	I	R
102	I	I
103	I	R
104	I	I
105	I	R
106	I	I
107	I	R
108	I	I
109	I	I
110	I	I
111	I	I
112	I	I
113	I	I
114	I	R
R(J,Q)	0	5
I(J,Q)	14	9
F (J,Q)	****	64.3
F (Q)		84.8
HYP 1=A	1	0

* * *

GROUP A
QUESTION 3

DOCUMENT NUMBERS

ANALYST	B400	B842	B260	B470	B100	B672	B430	B530	A430	A671	A610	A230	A582
101	I	R	I	R	R	R	R	R	I	R	R	I	I
102	I	I	I	I	I	I	I	I	I	I	I	I	I
103	I	I	I	I	I	I	I	I	I	I	I	I	R
104	I	I	I	I	I	I	I	I	I	I	I	I	I
105	I	I	I	I	I	R	I	R	I	I	I	I	R
106	I	I	I	R	I	I	R	I	I	I	I	I	R
107	I	I	R	I	I	I	R	I	I	I	I	R	R
108	I	I	I	I	I	I	I	I	I	I	I	I	I
109	I	I	R	I	I	I	I	I	I	I	I	I	R
110	I	I	I	I	I	I	I	I	I	I	I	I	I
111	I	I	I	I	I	I	I	I	I	I	I	I	I
112	I	I	I	I	I	I	I	I	I	I	I	I	R
113	I	I	R	R	I	I	R	I	I	I	I	I	I
114	I	I	I	I	I	I	I	I	I	I	I	I	R
R(J,Q)	0	1	3	3	1	2	4	2	0	1	1	1	7
I(J,Q)	14	13	11	11	13	12	10	12	14	13	13	13	7
F (J,Q)	****	92.9	78.6	78.6	92.9	85.7	71.4	85.7	****	92.9	92.9	92.9	50.0
HYP 1=A	1	1	1	1	1	1	0	1	1	1	1	1	0
			*				*				*		

GROUP A
QUESTION 3 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A600	A720	A800
101	R	I	I
102	I	I	I
103	I	R	I
104	I	I	I
105	R	R	I
106	I	I	I
107	I	I	I
108	I	I	I
109	I	R	I
110	I	I	I
111	I	I	I
112	I	R	I
113	I	I	I
114	I	I	I
R(J,Q)	2	4	0
I(J,Q)	12	10	14
F (J,Q)	85.7	71.4	****
F (Q)		85.7	
HYP 1=A	1	0	1

* * *

GROUP A
QUESTION A

DOCUMENT NUMBERS

ANALYST	B543	B750	B500	A100	B360	B200	B113	A370	A661	A650	B480	A290
101	R	I	I	I	R	I	I	I	R	R	I	R
102	I	I	I	I	I	I	I	I	I	I	I	I
103	I	I	I	I	R	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	R	I	I
105	I	I	I	I	I	R	I	I	I	I	I	I
106	R	I	I	I	R	I	R	I	R	R	I	I
107	I	I	I	I	R	I	I	I	I	R	I	I
108	I	I	I	I	R	I	I	I	I	I	I	I
109	I	I	I	I	I	I	I	I	I	I	I	I
110	I	I	I	I	I	I	I	I	I	I	I	I
111	I	I	I	I	I	I	I	I	I	I	I	I
112	I	I	I	I	I	I	I	I	I	I	I	I
113	I	I	I	I	I	I	I	I	I	R	I	I
114	I	I	I	I	R	I	I	I	I	I	I	I
R(J,Q)	2	0	0	0	6	1	1	0	2	5	0	1
I(J,Q)	12	14	14	14	8	13	13	14	12	9	14	13
F (J,Q)	85.7	****	****	****	57.1	92.9	92.9	****	85.7	64.3	****	92.9
F (Q)		89.3										
HYP 1=A	1	1	1	1	0	1	1	1	1	0	1	1
			*				*				*	

GROUP A
QUESTION 5

DOCUMENT NUMBERS

ANALYST	D110	D250	D700	D120	D170	D720	D400	D320	D270	D240	D880
101	I	R	R	R	R	I	R	R	I	R	R
102	I	I	R	I	R	I	I	R	I	I	R
103	I	R	R	R	I	I	I	R	I	I	R
104	I	I	R	I	I	I	I	R	I	R	R
105	I	R	R	R	R	I	R	R	I	R	R
106	I	R	R	R	I	I	I	R	I	I	R
107	R	R	R	I	R	I	R	R	I	R	R
108	I	I	R	I	R	I	I	R	I	R	R
109	I	R	R	R	R	I	R	R	R	I	R
110	I	I	R	R	I	I	R	R	I	R	R
111	I	I	R	R	R	I	R	R	I	R	R
112	I	R	R	R	I	I	R	R	I	R	R
113	R	I	R	R	R	I	R	R	I	R	R
114	I	R	R	R	I	I	R	R	I	R	R
R(J,Q)	2	8	14	10	8	0	9	14	1	10	14
I(J,Q)	12	6	0	4	6	14	5	0	13	4	0
F (J,Q)	85.7	57.1	****	71.4	57.1	****	64.3	****	92.9	71.4	****
F (Q)		81.8									
HYP 1=A	1	0	1	0	0	1	0	1	1	0	1
		*				*				*	

GROUP A
QUESTION 7

DOCUMENT NUMBERS

ANALYST	D110	D210	D130	D180	D230	D280	D310	D270	D300	B560	D370	D480	D400
101	R	R	I	R	R	R	I	I	R	I	I	R	R
102	I	I	I	I	I	I	I	I	I	I	I	I	I
103	R	I	I	I	I	I	I	I	R	I	I	I	R
104	I	I	I	I	I	I	I	I	I	I	I	I	I
105	I	I	I	I	I	I	I	I	R	I	I	I	R
106	R	R	I	R	R	R	R	R	R	R	I	R	R
107	I	R	I	I	R	R	I	I	I	I	I	R	R
108	R	I	I	I	I	I	I	I	I	I	I	I	R
109	R	I	I	I	R	I	I	I	R	I	I	R	R
110	I	I	I	I	I	I	I	I	R	I	I	I	R
111	I	I	I	I	I	I	I	I	I	I	I	I	R
112	I	I	I	I	R	R	R	I	I	I	I	I	R
113	R	I	I	I	R	I	I	I	I	I	I	I	R
114	I	I	I	I	R	I	I	I	I	I	I	I	R
R(J,Q)	6	3	0	2	7	4	2	1	6	1	0	4	12
I(J,Q)	8	11	14	12	7	10	12	13	8	13	14	10	2
F (J,Q)	57.1	78.6	****	85.7	50.0	71.4	85.7	92.9	57.1	92.9	****	71.4	85.7
HYP 1=A	0	1	1	1	0	0	1	1	0	1	1	0	1
			*				*				*		

GROUP A
QUESTION 7 - CONTINUED

DOCUMENT NUMBERS

ANALYST	D350	D410	D540	D620	D750	D640	D790	D630
101	I	R	R	I	R	I	I	R
102	I	I	I	I	R	I	I	I
103	I	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I
105	I	I	I	I	R	I	I	I
106	I	R	R	R	R	I	I	I
107	I	R	R	I	I	I	I	I
108	I	I	R	I	I	I	I	I
109	I	I	R	I	I	I	I	I
110	I	I	I	I	R	I	I	I
111	I	I	R	I	I	I	I	I
112	I	I	I	I	I	I	I	I
113	I	R	R	I	R	I	I	I
114	I	R	I	I	R	I	I	I
R(J,Q)	0	5	7	1	7	0	0	1
I(J,Q)	14	9	7	13	7	14	14	13
F (J,Q)	****	64.3	50.0	92.9	50.0	****	****	92.9
F (Q)		79.9						
HYP 1-A	1	0	0	1	0	1	1	1
		*				*		*

GROUP A
QUESTION 8

DOCUMENT NUMBERS

ANALYST	D780	D600	D530	D810	D280	D150	D440	D610	D660	D830	D360	D820	D620
101	R	R	I	R	R	R	R	R	R	R	I	R	R
102	I	I	I	R	I	R	R	I	R	R	I	I	I
103	I	I	I	R	I	R	I	I	I	I	I	I	I
104	I	I	I	R	I	R	R	I	I	I	I	I	I
105	I	R	I	R	I	R	R	I	I	R	I	I	I
106	R	R	I	R	I	R	R	R	I	R	I	R	I
107	I	R	I	R	I	R	R	I	I	I	I	I	I
108	I	I	I	R	R	R	R	I	I	I	I	I	I
109	I	I	I	I	I	R	R	I	I	I	I	I	I
110	I	I	I	R	I	R	R	I	I	I	I	I	I
111	I	R	I	R	R	R	R	I	I	I	I	I	I
112	I	R	I	R	I	R	R	I	I	I	I	I	I
113	I	I	I	R	R	R	R	I	I	I	I	I	I
114	I	I	I	R	R	R	R	I	I	I	I	I	I
R(J,Q)	2	6	0	13	5	14	13	2	2	4	0	2	1
I(J,Q)	12	8	14	1	9	0	1	12	12	10	14	12	13
F (J,Q)	85.7	57.1	****	92.9	64.3	****	92.9	85.7	85.7	71.4	****	85.7	92.9
F (Q)		85.7											
HYP 1-A	1	0	1	1	0	1	1	1	1	0	1	1	1
			*			*					*		

GROUP A
QUESTION 9

DOCUMENT NUMBERS

ANALYST	A980	B780	B650	B170	B900	B252	B700	A880	A190	A710	A750	A690
101	R	I	I	I	I	I	I	R	I	I	I	I
102	I	I	I	I	I	I	I	R	I	I	I	I
103	I	I	I	I	I	I	I	R	I	I	I	R
104	I	I	I	I	I	I	I	R	I	I	I	I
105	R	R	I	I	R	I	I	R	I	I	I	I
106	I	I	I	I	R	I	I	R	I	I	R	R
107	I	I	I	I	I	I	I	R	I	I	I	I
108	I	I	I	I	I	I	I	R	I	I	I	I
109	I	I	I	I	I	I	I	I	I	I	I	I
110	I	I	I	I	I	I	I	R	I	I	I	I
111	I	I	I	I	I	I	I	R	I	I	I	R
112	I	I	I	I	R	I	I	R	I	I	I	I
113	I	I	I	I	I	I	I	R	I	I	I	I
114	I	I	I	I	I	I	I	R	I	I	I	R
R(J,Q)	2	1	0	0	3	0	0	13	0	0	1	4
I(J,Q)	12	13	14	14	11	14	14	1	14	14	13	10
F (J,Q)	85.7	92.9	****	****	78.6	****	****	92.9	****	****	92.9	71.4
F (Q)		92.9										
HYP 1=A	1	1	1	1	1	1	1	1	1	1	1	0
			*			*					*	

GROUP A
QUESTION 10

DOCUMENT NUMBERS

ANALYST	B700	B550	B960	B500	B520	B610	B540	B160	B350	B360	B570	A110	A200
101	R	R	I	I	R	R	I	R	I	I	I	R	R
102	I	I	I	I	I	I	I	I	I	I	I	I	I
103	I	I	I	I	R	I	I	I	I	I	I	R	I
104	I	I	I	I	I	I	I	I	I	I	I	R	I
105	I	R	I	I	R	I	I	I	I	I	I	R	I
106	I	I	I	I	I	I	I	R	I	I	I	I	I
107	I	R	I	I	I	I	I	I	I	I	I	I	I
108	I	I	I	I	I	I	I	I	I	I	I	I	I
109	I	R	I	I	I	R	I	I	I	I	I	I	R
110	I	I	I	I	I	I	I	I	I	I	I	I	I
111	I	I	I	I	I	I	I	I	I	I	I	I	I
112	I	I	I	I	I	I	I	I	I	I	I	I	I
113	I	R	I	I	I	I	I	I	I	I	I	I	I
114	I	I	I	I	I	I	I	I	I	I	I	R	I
R(J,Q)	1	5	0	0	3	2	0	2	0	0	0	5	2
I(J,Q)	13	9	14	14	11	12	14	12	14	14	14	9	12
F (J,Q)	92.9	64.3	****	****	78.6	85.7	****	85.7	****	****	****	64.3	85.7
HYP 1=A	1	0	1	1	1	1	1	1	1	1	1	0	1
			*				*				*		

GROUP A
QUESTION 10 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A000	A750	B290	A700	A350	A590	A900	A630	A672	A180
101	I	I	R	I	I	R	R	I	I	I
102	I	I	I	I	I	I	I	I	I	I
103	I	I	I	I	R	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	I
105	I	I	I	I	I	I	I	I	I	I
106	I	I	I	I	R	R	I	I	I	I
107	I	I	I	I	I	I	R	I	I	I
108	I	I	I	I	I	I	I	I	I	I
109	I	I	I	I	I	I	R	I	I	I
110	I	I	I	I	I	I	I	I	I	I
111	I	I	I	I	I	I	I	I	I	I
112	I	I	I	I	I	I	I	I	I	I
113	I	I	I	I	I	I	R	I	I	I
114	I	I	I	I	I	I	R	I	I	I
R(J,Q)	0	0	1	0	2	2	5	0	0	0
I(J,Q)	14	14	13	14	12	12	9	14	14	14
F (J,Q)	****	****	92.9	****	85.7	85.7	64.3	****	****	****
F (Q)		90.7								
HYP 1=A	1	1	1	1	1	1	0	1	1	1
			*			*				*

GROUP A
QUESTION 11

DOCUMENT NUMBERS

ANALYST	A320	A600	A870	A110	A200	A930	A942	A770	A280	A400	B170	B160	B710
101	I	R	R	R	R	I	R	I	R	R	I	R	R
102	I	I	I	I	I	I	I	I	I	I	I	I	I
103	I	I	R	I	I	I	I	I	I	I	I	I	I
104	I	I	R	I	I	I	I	I	I	I	I	I	I
105	I	I	R	I	I	I	R	I	I	I	I	I	I
106	I	I	R	R	I	I	I	I	R	R	I	R	I
107	I	I	R	I	I	I	R	I	I	I	I	R	I
108	I	I	R	I	I	I	I	I	I	I	I	I	I
109	I	I	R	R	I	I	R	I	I	I	I	R	I
110	I	I	R	I	I	I	I	I	I	I	I	R	I
111	I	I	R	I	I	I	I	I	I	I	I	I	I
112	I	I	R	I	I	I	I	I	I	R	I	I	I
113	I	I	R	R	I	I	R	I	I	I	I	I	I
114	I	I	R	R	I	I	I	I	R	R	I	I	I
R(J,Q)	0	1	13	5	1	0	5	0	3	4	0	5	1
I(J,Q)	14	13	1	9	13	14	9	14	11	10	14	9	13
F (J,Q)	****	92.9	92.9	64.3	92.9	****	64.3	****	78.6	71.4	****	64.3	92.9
HYP 1=A	1	1	1	0	1	1	0	1	1	0	1	0	1
			*				*				*		

GROUP A
QUESTION 11 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A250	B292	B470	B841	B691	B852	B901	B850	A521	B700	B692
101	I	R	R	I	I	R	I	R	I	R	I
102	I	I	I	I	I	I	I	I	I	I	I
103	I	I	I	I	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	R	I
105	I	I	R	I	I	I	I	R	I	R	I
106	I	I	R	I	I	I	I	R	R	R	I
107	I	I	R	I	I	I	I	I	I	R	I
108	I	I	I	I	I	I	I	I	I	R	I
109	I	I	R	I	I	I	I	I	R	R	I
110	I	I	I	I	I	I	I	I	I	R	I
111	I	I	I	I	I	I	I	I	I	R	I
112	I	I	I	I	I	I	I	I	I	R	I
113	I	I	I	I	I	I	I	I	I	I	I
114	I	I	I	I	I	I	I	I	R	I	I
R(J,Q)	0	1	5	0	0	1	0	3	3	10	0
I(J,Q)	14	13	9	14	14	13	14	11	11	4	14
F (J,Q)	****	92.9	64.3	****	****	92.9	****	78.6	78.6	71.4	****
F (Q)		87.2									
HYP 1=A	1	1	0	1	1	1	1	1	1	0	1
		*				*				*	

GROUP A
QUESTION 13

DOCUMENT NUMBERS

ANALYST	B490	A600	A420	B421	A810	B600	A900	A730	B240	A662	B610	A530	A360
101	R	I	I	I	R	I	R	I	R	R	I	I	I
102	I	I	I	I	R	I	I	I	I	I	I	I	I
103	I	I	R	I	I	I	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	R	I	I	I
105	R	I	I	I	R	I	I	I	I	R	I	I	I
106	R	I	I	R	R	I	I	I	I	R	I	I	I
107	I	I	I	I	I	I	I	I	I	I	I	I	I
108	I	I	I	I	R	I	I	I	I	I	I	I	I
109	R	I	I	I	R	I	I	I	I	R	I	I	I
110	I	I	I	I	I	I	I	I	R	I	I	I	I
111	I	I	I	I	R	I	I	I	I	R	I	I	I
112	R	I	I	I	R	I	I	I	R	R	I	I	I
113	R	I	I	I	R	I	I	I	I	R	I	I	I
114	R	I	I	I	R	I	I	I	R	I	I	I	I
R(J,Q)	7	0	1	1	10	0	1	0	4	8	0	0	0
I(J,Q)	7	14	13	13	4	14	13	14	10	6	14	14	14
F (J,Q)	50.0	****	92.9	92.9	71.4	****	92.9	****	71.4	57.1	****	****	****
HYP 1=A	0	1	1	1	0	1	1	1	0	0	1	1	1
			*				*				*		

GROUP A
QUESTION 13 - CONTINUED

DOCUMENT NUMBERS

ANALYST	B620	A212	A990	B902	A310	A450	B730
101	I	R	I	R	R	I	I
102	I	I	I	I	I	I	I
103	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I
105	I	I	I	R	I	I	I
106	I	R	I	R	I	I	I
107	I	I	I	R	I	I	I
108	I	I	I	I	I	I	I
109	I	R	I	R	I	I	I
110	I	I	I	I	I	I	I
111	I	I	I	I	I	I	I
112	I	I	I	R	I	I	I
113	I	I	I	R	I	I	I
114	I	I	I	I	I	I	I
R(J,Q)	0	3	0	7	1	0	0
I(J,Q)	14	11	14	7	13	14	14
F (J,Q)	****	78.6	****	50.0	92.9	****	****
F (Q)		87.5					
HYP 1=A	1	1	1	0	1	1	1
		*			*		*

GROUP A
QUESTION 15

DOCUMENT NUMBERS

ANALYST	B620	B380	B940	A620	A140	A500	A900	A250	A820	A120
101	R	I	I	I	I	I	I	I	I	I
102	I	I	I	I	R	I	I	I	I	I
103	I	I	I	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	I
105	R	I	I	I	I	I	I	I	I	I
106	R	I	I	R	R	I	I	I	I	I
107	R	I	I	I	I	I	I	I	I	I
108	R	I	I	I	I	I	I	I	I	I
109	R	I	I	R	I	I	I	I	I	I
110	R	I	I	I	I	I	I	I	I	I
111	R	I	I	I	I	I	I	I	I	I
112	R	I	I	I	I	I	I	I	I	I
113	R	I	I	I	I	I	I	I	I	I
114	R	I	I	I	I	I	I	I	I	I
R(J,Q)	11	0	0	2	2	0	0	0	0	0
I(J,Q)	3	14	14	12	12	14	14	14	14	14
F (J,Q)	78.6	****	****	85.7	85.7	****	****	****	****	****
F (Q)		95.0								
HYP 1=A	1	1	1	1	1	1	1	1	1	1
		*				*			*	

GROUP A
QUESTION 16

DOCUMENT NUMBERS

ANALYST	B310	B860	B490	A950	B400	B890	B280	A760	B390	A930	A020	A790	B540
101	I	I	R	I	I	I	I	I	R	R	R	I	R
102	I	I	I	I	I	I	I	I	I	I	I	I	I
103	I	I	I	I	I	I	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	I	I	I	I
105	I	R	R	R	I	I	I	I	R	I	I	I	R
106	I	I	R	R	I	I	I	I	R	I	I	I	I
107	I	I	R	I	I	I	I	I	I	I	I	I	I
108	I	I	I	I	I	I	I	I	I	I	I	I	I
109	I	I	I	I	I	I	I	I	I	I	I	I	I
110	I	I	I	I	I	I	I	I	I	I	I	I	I
111	I	I	I	I	I	I	I	I	I	I	I	I	I
112	I	I	I	I	I	I	I	I	I	I	I	I	I
113	I	I	I	I	I	I	I	I	I	I	I	I	R
114	I	I	I	I	I	I	I	I	I	I	I	I	I
R(J,Q)	0	1	4	2	0	0	0	0	3	1	1	0	3
I(J,Q)	14	13	10	12	14	14	14	14	11	13	13	14	11
F (J,Q)	****	92.9	71.4	85.7	****	****	****	****	78.6	92.9	92.9	****	78.6
F		87.6											
HYP 1=A	1	1	0	1	1	1	1	1	1	1	1	1	1
			*				*				*		

GROUP A
QUESTION 16 - CONTINUED

DOCUMENT NUMBERS

ANALYST	B693	B382	B671	B980	B920	A740	A520	A260	A540	B950	B180	A602
101	I	R	I	I	I	R	I	I	R	R	I	I
102	I	I	I	I	I	I	I	I	I	R	I	I
103	I	I	I	I	I	I	I	I	I	I	I	I
104	I	I	I	I	I	I	I	I	I	R	I	I
105	I	R	I	I	I	R	I	I	R	R	I	I
106	I	I	I	I	I	I	I	R	I	R	I	I
107	I	R	I	I	I	R	I	I	R	R	I	I
108	I	I	I	I	I	I	I	I	I	R	I	I
109	I	R	I	I	I	R	I	I	I	R	I	I
110	I	R	I	I	I	I	I	I	I	R	I	I
111	I	I	I	I	I	I	I	I	I	R	I	I
112	I	I	I	I	I	R	I	I	I	R	I	I
113	I	R	I	I	I	I	I	I	I	R	I	I
114	I	I	I	I	I	I	I	I	I	R	I	I
R(J,Q)	0	6	0	0	0	5	0	1	3	13	0	0
I(J,Q)	14	8	14	14	14	9	14	13	11	1	14	14
F (J,Q)	****	57.1	****	****	****	64.3	****	92.9	78.6	92.9	****	****
F (Q)		91.1										
F		87.5										
HYP 1=A	1	0	1	1	1	0	1	1	1	1	1	1
		*				*				*		

GROUP 8
QUESTION 2

DOCUMENT NUMBERS

ANALYST	B420	B910	B200	B460	B220	B580	B600	A580	A410	A550	B770	A570	A480
201	I	I	I	I	I	I	I	R	I	I	R	I	I
203	I	I	I	I	I	I	I	I	I	I	R	I	I
204	I	I	I	I	I	I	I	I	I	I	I	I	I
205	I	R	I	R	I	I	I	R	R	I	I	I	I
206	I	R	I	I	I	I	I	I	I	I	I	I	I
207	I	I	I	R	I	I	I	R	I	I	R	I	I
208	I	I	R	R	I	I	R	I	R	I	R	I	I
210	I	I	I	I	I	I	I	I	I	I	I	I	I
211	I	I	I	I	R	I	I	I	I	I	I	I	I
212	I	R	I	I	I	I	I	I	I	I	I	I	I
213	I	I	I	I	I	I	R	I	I	I	I	I	R
214	R	I	I	I	I	I	I	I	I	I	R	I	I
215	I	I	I	I	I	I	R	I	I	I	I	I	I
216	I	I	I	I	I	I	I	I	I	I	R	I	I
R(J,Q)	1	3	1	3	1	0	3	3	2	0	6	0	1
I(J,Q)	13	11	13	11	13	14	11	11	12	14	8	14	13
F (J,Q)	92.9	78.6	92.9	78.6	92.9	****	78.6	78.6	85.7	****	57.1	****	92.9
HYP 1=A	1	1	1	1	1	1	1	1	1	1	0	1	1
			*				*				*		

GROUP B
QUESTION 2 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A210	A970
201	I	R
203	I	I
204	I	I
205	I	I
206	I	I
207	R	I
208	R	R
210	I	I
211	I	I
212	I	I
213	I	I
214	I	I
215	I	R
216	I	I
R(J,Q)	2	3
I(J,Q)	12	11
F (J,Q)	85.7	78.6
F (Q)		86.2
HYP 1=A	1	1

* * *

GROUP 8
QUESTION 3

DOCUMENT NUMBERS

ANALYST	B400	B842	B260	B470	B100	B672	B430	B530	A430	A671	A610	A230	A582
201	I	I	I	I	I	I	I	I	I	I	I	I	I
203	I	I	I	I	R	R	I	R	R	I	I	I	R
204	I	I	I	R	R	R	I	R	R	I	I	I	R
205	I	I	I	R	I	R	I	R	R	I	I	I	R
206	I	I	I	I	I	R	R	I	I	I	I	I	R
207	I	I	I	I	I	I	I	I	R	I	I	I	R
208	R	I	I	R	R	R	R	R	R	I	I	I	R
210	I	I	I	I	I	R	I	I	R	I	I	I	I
211	I	I	I	R	R	R	I	R	R	I	I	R	R
212	R	I	I	R	R	R	R	R	I	I	I	I	R
213	R	I	I	I	I	I	I	R	I	I	I	I	R
214	I	I	I	I	I	R	I	R	R	I	I	I	R
215	I	I	I	I	I	I	R	I	R	I	I	I	R
216	I	I	I	I	I	I	I	I	I	I	I	I	I
R(J,Q)	3	0	0	5	5	9	4	8	9	0	0	1	11
I(J,Q)	11	14	14	9	9	5	10	6	5	14	14	13	3
F (J,Q)	78.6	****	****	64.3	64.3	64.3	71.4	57.1	64.3	****	****	92.9	78.6
HYP 1=A	1	1	1	0	0	0	0	0	0	1	1	1	1
			*				*				*		

GROUP B
QUESTION 3 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A600	A720	A800
201	I	I	I
203	I	R	I
204	I	R	I
205	I	R	I
206	I	R	I
207	I	I	I
208	I	R	R
210	I	R	I
211	I	R	I
212	I	R	I
213	I	R	I
214	I	R	I
215	I	R	I
216	I	I	I
R(J,Q)	0	11	1
I(J,Q)	14	3	13
F (J,Q)	****	78.6	92.9
F (Q)		81.7	
HYP 1=A	1	1	1

* * *

GROUP 8
QUESTION 4

DOCUMENT NUMBERS

ANALYST	B543	B750	B500	A100	B360	B200	B113	A370	A661	A650	A480	A290
201	I	I	I	I	I	I	I	I	I	I	I	I
203	I	I	I	I	I	I	I	I	I	I	I	I
204	I	I	I	I	I	I	I	I	R	I	I	I
205	I	I	I	I	I	I	I	I	R	R	I	I
206	I	I	I	I	I	I	I	I	I	R	I	I
207	R	I	I	I	I	I	I	I	R	R	I	I
208	R	R	I	I	R	I	R	I	R	R	I	I
210	R	I	I	I	I	I	I	I	R	R	I	I
211	I	I	I	I	R	R	R	I	I	I	I	I
212	I	R	I	I	R	I	I	R	R	R	I	I
213	I	I	I	I	I	I	I	I	I	I	I	I
214	I	I	I	I	R	I	I	I	R	I	I	I
215	I	I	I	I	I	I	I	I	I	R	I	I
216	I	I	I	I	I	I	I	I	R	R	I	I
R(J,Q)	3	2	0	0	4	1	2	1	8	8	0	0
I(J,Q)	11	12	14	14	10	13	12	13	6	6	14	14
F (J,Q)	78.6	85.7	****	****	71.4	92.9	85.7	92.9	57.1	57.1	****	****
F (Q)		85.1										
HYP 1=A	1	1	1	1	0	1	1	1	0	0	1	1
			*			*				*		

GROUP B
QUESTION 5

DOCUMENT NUMBERS

ANALYST	D110	D250	D700	D120	D170	D720	D400	D320	D270	D240	D880	
201	R	R	R	R	R	R	R	R	R	R	R	
203	R	I	R	I	I	R	R	R	I	R	R	
204	R	R	R	I	R	R	R	R	R	R	R	
205	R	R	R	I	I	I	R	R	I	I	R	
206	I	I	R	R	R	R	I	R	I	R	R	
207	R	I	I	R	R	I	I	R	I	R	R	
208	R	R	R	R	R	R	R	R	I	R	R	
210	R	R	R	R	R	R	R	R	R	R	R	
211	I	R	R	I	R	I	R	I	R	R	R	
212	R	R	R	I	R	R	R	R	R	R	R	
213	R	R	R	R	R	R	R	R	I	R	R	
214	I	R	R	I	I	I	R	R	I	R	R	
215	I	I	R	R	R	I	R	R	I	R	R	
216	R	I	I	R	R	I	I	R	I	I	R	
R(J,Q)	10	9	12	8	11	8	11	13	5	12	14	
I(J,Q)	4	5	2	6	3	6	3	1	9	2	0	
F (J,Q)	71.4	64.3	85.7	57.1	78.6	57.1	78.6	92.9	64.3	85.7	****	
F (Q)		76.0										
HYP 1=A	0	0	1	0	1	0	1	1	0	1	1	
			*				*			*		

GROUP B
QUESTION 7

DOCUMENT NUMBERS

ANALYST	D110	D210	D130	D180	D230	D280	D310	D270	D300	D560	D370	D480	D400
201	R	R	I	R	I	I	I	I	R	R	I	I	I
203	R	I	R	R	I	I	I	R	R	I	R	R	R
204	R	R	I	I	R	I	I	I	R	I	R	R	I
205	R	R	R	R	R	R	I	I	I	I	I	R	I
206	R	I	R	R	R	I	I	I	R	R	I	R	R
207	R	R	I	R	R	I	I	I	I	R	I	I	R
208	R	R	R	R	R	R	R	R	R	I	R	R	R
210	I	R	R	R	R	R	I	R	R	R	R	R	R
211	R	I	I	I	I	R	I	I	R	I	I	I	R
212	R	R	R	R	R	R	R	R	R	R	R	R	R
213	R	R	R	R	R	R	R	I	R	I	R	R	R
214	R	R	R	R	I	R	R	R	R	I	R	R	R
215	R	R	R	I	I	I	I	I	R	I	I	R	R
216	R	I	R	R	I	R	R	I	R	I	I	I	I
R(J,Q)	13	10	10	11	8	8	5	5	12	5	7	10	10
I(J,Q)	1	4	4	3	6	6	9	9	2	9	7	4	4
F (J,Q)	92,9	71.4	71.4	78,6	57.1	57.1	64.3	64.3	85.7	64.3	50.0	71.4	71.4
HYP 1=A	1	0	0	1	0	0	0	0	1	0	0	0	0
			*			*					*		

GROUP B
QUESTION 7 - CONTINUED

DOCUMENT NUMBERS

ANALYST	D350	D410	D540	D620	D750	D640	D790	D630
201	I	I	I	I	I	I	I	I
203	R	I	I	I	I	I	I	R
204	I	I	I	I	R	R	I	R
205	I	I	I	I	I	I	I	I
206	I	I	I	I	R	I	R	R
207	R	I	I	I	R	I	I	I
208	R	R	R	I	R	R	I	R
210	R	R	R	R	R	I	R	R
211	I	I	I	I	R	I	I	R
212	R	R	R	R	R	I	I	R
213	R	R	R	R	R	R	I	R
214	R	R	R	R	R	R	I	R
215	R	I	I	I	I	I	I	I
216	R	R	I	I	R	I	I	R
R(J,Q)	9	6	5	4	10	4	2	10
I(J,Q)	5	8	9	10	4	10	12	4
F (J,Q)	64.3	57.1	64.3	71.4	71.4	71.4	85.7	71.4
F (Q)		69.4						
HYP 1=A	0	0	0	0	0	0	1	0
		*				*		*

GROUP B
QUESTION 8

DOCUMENT NUMBERS

ANALYST	D780	D600	D530	D810	D280	D150	D440	D610	D660	D830	D360	D820	D620
201	R	R	R	R	R	R	R	R	R	R	R	R	R
203	I	R	I	I	I	R	R	I	I	I	I	I	I
204	I	R	I	I	I	R	R	I	I	R	I	I	I
205	I	R	I	I	R	R	R	I	I	I	I	I	I
206	I	I	I	I	I	R	I	I	I	R	I	I	I
207	I	R	I	R	I	I	R	I	I	R	I	R	I
208	R	R	I	R	R	R	R	I	R	I	I	I	R
210	I	I	I	R	I	R	R	I	I	I	I	I	I
211	R	I	I	R	I	R	R	I	I	I	I	R	I
212	I	R	I	I	I	R	R	I	I	I	I	I	I
213	R	R	I	I	R	R	R	I	I	R	I	I	I
214	I	R	I	R	I	I	I	R	R	I	I	R	I
215	I	I	I	R	I	R	R	I	R	I	I	R	I
216	I	I	I	I	R	R	R	I	I	R	I	I	R
R(J,Q)	4	9	1	7	5	12	12	2	4	6	1	5	3
I(J,Q)	10	5	13	7	9	2	2	12	10	8	13	9	11
F (J,Q)	71.4	64.3	92.9	50.0	64.3	85.7	85.7	85.7	71.4	57.1	92.9	64.3	78.6
F (Q)		74.2											
HYP 1=4	0	0	1	0	0	1	1	1	0	0	1	0	1
			*				*				*		

GROUP 8
QUESTION 9

DOCUMENT NUMBERS

ANALYST	A980	B780	B650	B170	B900	B252	B700	A880	A190	A710	A750	A690
201	I	I	I	I	I	I	I	R	I	I	I	R
203	I	I	I	I	I	R	I	I	I	I	R	I
204	I	I	I	I	I	I	I	R	I	I	I	I
205	I	I	I	I	I	I	I	I	I	I	I	R
206	I	I	I	I	I	I	I	R	I	I	R	R
207	I	I	I	I	I	I	I	R	I	I	I	R
208	R	I	I	I	I	I	I	R	I	I	R	R
210	I	I	I	I	I	I	I	R	I	I	I	I
211	I	R	I	I	R	I	I	R	I	I	R	I
212	I	R	I	I	I	I	I	R	I	I	I	R
213	I	I	I	I	I	I	I	I	I	I	I	I
214	I	I	I	I	I	I	I	I	I	I	I	R
215	I	I	I	I	I	I	I	R	I	I	I	R
216	I	I	I	I	I	I	I	R	I	I	I	R
R(J,Q)	1	2	0	0	1	1	0	10	0	0	4	9
I(J,Q)	13	12	14	14	13	13	14	4	14	14	10	5
F (J,Q)	92.9	85.7	****	****	92.9	92.9	****	71.4	****	****	71.4	64.3
F (Q)		89.3										
HYP 1=A	1	1	1	1	1	1	1	0	1	1	0	0
			*			*				*		

GROUP B
QUESTION 10

DOCUMENT NUMBERS

ANALYST	B700	B550	B960	B500	B520	B610	B540	B160	B350	B360	B570	A110	A200
201	R	I	I	I	R	I	I	I	I	I	I	I	I
203	I	R	I	I	I	I	I	I	I	I	I	I	I
204	I	I	I	I	I	R	I	I	I	R	I	I	I
205	I	I	I	I	I	I	I	I	I	I	I	I	I
206	I	I	I	I	I	I	I	I	I	I	I	I	I
207	I	I	I	I	I	R	R	I	I	I	I	I	I
208	R	R	R	R	R	R	R	R	I	R	R	I	R
210	I	R	I	I	I	R	I	I	I	I	I	I	R
211	R	R	I	I	I	I	I	I	I	I	I	I	I
212	R	R	I	I	I	I	I	I	I	I	I	I	I
213	R	I	I	I	I	R	R	I	I	R	I	I	I
214	I	R	I	I	I	I	I	I	I	I	I	R	I
215	R	R	I	I	R	I	I	I	I	I	I	I	I
216	I	I	I	I	I	I	I	I	I	I	I	I	I
R(J,Q)	6	7	1	1	3	5	3	1	0	3	1	1	2
I(J,Q)	8	7	13	13	11	9	11	13	14	11	13	13	12
F (J,Q)	57.1	50.0	92.9	92.9	78.6	64.3	78.6	92.9	****	78.6	92.9	92.9	85.7
HYP 1=A	0	0	1	1	1	0	1	1	1	1	1	1	1
			*				*				*		

GROUP 9
QUESTION 10 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A000	A750	B290	A700	A350	A590	A900	A630	A672	A180
201	I	I	I	I	R	I	I	I	I	I
203	I	I	I	I	I	I	R	I	I	I
204	I	I	I	I	I	I	I	I	I	I
205	I	I	I	I	I	I	I	I	I	I
206	I	I	I	I	I	I	R	I	I	I
207	I	I	I	I	R	I	R	I	I	I
208	I	R	R	R	I	R	R	R	I	R
210	I	I	I	I	I	I	R	I	I	I
211	I	I	I	I	I	I	R	I	I	I
212	I	I	I	I	I	I	I	I	I	I
213	I	I	I	I	I	I	I	I	I	I
214	I	I	I	I	I	I	R	I	I	I
215	I	I	I	I	I	I	R	I	I	I
216	I	I	I	I	I	I	R	I	I	I
R(J,Q)	0	1	1	1	2	1	9	1	0	1
I(J,Q)	14	13	13	13	12	13	5	13	14	13
F (J,Q)	****	92.9	92.9	92.9	85.7	92.9	64.3	92.9	****	92.9
F (Q)		85.4								
HYP 1=A	1	1	1	1	1	1	0	1	1	1
			*				*			*

GROUP 8
QUESTION 11

DOCUMENT NUMBERS

ANALYST	A320	A600	A870	A110	A200	A930	B942	A770	A280	A400	B170	B160	B710
201	I	I	I	I	I	I	I	I	I	I	I	I	I
203	I	I	I	R	I	I	R	I	I	I	I	R	I
204	I	I	I	I	I	I	R	I	I	I	I	R	I
205	I	I	R	I	I	I	I	I	I	I	I	R	I
206	I	I	I	I	I	I	R	I	I	I	I	I	I
207	I	I	R	I	I	I	I	I	I	I	I	R	I
208	I	I	R	R	R	I	R	I	I	R	I	R	I
210	I	I	I	I	I	I	R	I	I	I	I	I	I
211	I	I	I	I	I	I	R	I	I	I	I	I	I
212	I	I	R	I	I	I	R	I	I	I	I	I	I
213	I	I	R	R	I	I	I	I	R	R	I	I	I
214	I	I	R	I	I	I	I	I	I	I	I	I	I
215	I	I	R	I	I	I	R	I	I	I	I	R	I
216	I	I	R	I	I	I	I	I	R	I	I	R	I
R(J,Q)	0	0	8	3	1	0	8	0	2	2	0	7	0
I(J,Q)	14	14	6	11	13	14	6	14	12	12	14	7	14
F(J,Q)	****	****	57.1	78.6	92.9	****	57.1	****	85.7	85.7	****	50.0	****
HYP 1=A	1	1	0	1	1	1	0	1	1	1	1	0	1
			*				*				*		

GROUP 8
QUESTION 11 - CONTINUED

DOCUMENT NUMBERS

ANALYST	A250	B292	B470	B841	B691	B852	B901	B850	A521	B700	B692
201	I	I	I	I	I	I	I	I	I	I	I
203	I	I	R	I	I	I	I	I	I	R	I
204	I	I	I	I	I	I	I	I	I	R	I
205	I	I	I	I	I	I	I	I	R	R	I
206	I	I	I	I	I	I	I	I	I	I	I
207	I	I	I	I	I	I	I	I	R	R	I
208	I	R	R	R	I	R	R	R	R	R	I
210	I	I	R	I	I	I	I	I	I	R	I
211	I	I	I	I	I	I	I	I	I	R	I
212	I	I	I	I	I	I	I	I	I	R	I
213	I	I	R	I	I	I	I	I	I	R	I
214	I	I	I	I	I	I	R	I	R	R	I
215	I	I	I	I	I	I	I	I	R	R	I
216	I	I	I	I	I	I	I	I	I	R	I
R(J,Q)	0	1	4	1	0	1	2	1	5	12	0
I(J,Q)	14	13	10	13	14	13	12	13	9	2	14
F (J,Q)	****	92.9	71.4	92.9	****	92.9	85.7	92.9	64.3	85.7	****
F (Q)		86.9									
HYP 1=A	1	1	0	1	1	1	1	1	0	1	1
			*				*			*	

GROUP B
QUESTION 13

DOCUMENT NUMBERS

ANALYST	B490	A600	A420	B421	A810	B600	A900	A730	B240	A662	B610	A530	A360
201	I	I	I	I	R	I	I	I	I	I	I	I	I
203	R	I	I	I	R	I	I	I	I	R	I	I	I
204	R	I	I	I	R	I	I	I	I	R	I	I	I
205	I	I	I	I	I	I	I	I	I	I	I	I	I
206	I	I	I	I	I	I	I	I	R	I	I	I	I
207	I	I	I	I	I	I	I	I	I	R	I	R	I
208	R	I	I	I	R	I	R	I	I	I	I	I	I
210	R	I	I	I	R	I	I	I	I	R	I	I	I
211	I	I	I	R	R	I	I	I	R	R	I	I	I
212	R	I	I	I	R	I	I	I	I	R	I	I	I
213	R	R	I	I	R	I	R	I	I	R	I	I	I
214	I	I	I	I	R	I	I	I	I	R	I	I	I
215	I	I	I	I	R	I	I	I	I	R	I	I	I
216	I	I	I	I	I	I	I	I	I	I	I	I	I
R(J,Q)	6	1	0	1	10	0	2	0	2	9	0	1	0
I(J,Q)	8	13	14	13	4	14	12	14	12	5	14	13	14
F (J,Q)	57.1	92.9	****	92.9	71.4	****	85.7	****	85.7	64.3	****	92.9	****
HYP 1=A	0	1	1	1	0	1	1	1	1	0	1	1	1
			*				*				*		

GROUP B
QUESTION 13 - CONTINUED

DOCUMENT NUMBERS

ANALYST	B620	A212	A990	B902	A310	A450	B730
201	I	I	I	I	I	I	I
203	I	I	I	R	I	I	I
204	I	R	I	R	I	I	I
205	I	I	I	I	I	I	I
206	I	I	I	R	I	I	I
207	I	I	I	R	I	I	I
208	R	R	I	I	I	I	R
210	I	I	I	R	I	I	I
211	I	I	I	R	R	I	I
212	I	I	I	I	I	I	I
213	I	I	I	I	I	I	I
214	I	I	I	R	I	I	I
215	I	I	I	R	I	I	I
216	I	I	I	I	I	I	I
R(J,Q)	1	2	0	8	1	0	1
I(J,Q)	13	12	14	6	13	14	13
F (J,Q)	92,9	85,7	****	57,1	92,9	****	92,9
F (Q)		88,2					
HYP 1=A	1	1	1	0	1	1	1
		*			*		*

GROUP B
QUESTION 15

DOCUMENT NUMBERS

ANALYST	B620	B380	B940	A620	A140	A500	A900	A250	A820	A120
201	R	I	I	I	R	I	I	I	I	I
203	I	I	I	I	I	I	I	R	I	I
204	R	I	I	R	I	I	I	I	I	I
205	I	I	I	I	I	I	I	I	I	I
206	R	I	I	I	I	I	I	I	I	I
207	I	I	I	I	I	I	I	R	I	I
208	R	I	I	R	I	I	R	I	I	I
210	R	I	I	I	I	I	I	I	I	I
211	R	I	I	R	R	I	I	I	I	I
212	I	I	I	I	I	I	I	I	I	I
213	I	I	I	R	I	I	I	R	I	I
214	I	I	I	R	I	I	I	I	I	I
215	R	I	I	I	I	I	I	I	I	I
216	I	I	I	I	I	I	I	I	I	I
R(J,Q)	7	0	0	5	2	0	1	3	0	0
I(J,Q)	7	14	14	9	12	14	13	11	14	14
F (J,Q)	50.0	****	****	64.3	85.7	****	92.9	78.6	****	****
F (Q)		87.1								
HYP 1=A	0	1	1	0	1	1	1	1	1	1
		*				*			*	

GROUP 8
QUESTION 16

DOCUMENT NUMBERS

ANALYST	B310	B860	B490	A950	B400	B890	B280	A760	B390	A930	A020	A790	B540
201	I	I	R	R	I	I	I	I	R	R	I	R	R
203	I	R	R	R	I	I	I	I	R	R	I	I	I
204	I	I	I	R	I	I	I	I	R	I	R	I	I
205	I	I	I	I	I	I	I	I	I	I	I	I	I
206	I	I	I	R	I	I	I	R	R	I	I	R	I
207	I	I	I	R	I	I	I	I	R	R	I	R	R
208	I	R	R	R	I	I	R	R	R	R	R	I	R
210	I	I	I	R	I	I	I	I	R	I	I	I	R
211	I	R	R	R	I	I	I	I	R	R	R	R	R
212	I	R	R	R	I	I	I	I	R	R	I	R	R
213	I	I	I	I	I	I	I	I	R	R	I	I	I
214	I	I	I	I	I	I	I	I	I	I	I	I	I
215	I	I	I	I	I	I	I	I	R	R	I	I	I
216	I	I	I	I	I	I	I	I	I	I	I	I	I
R(J,Q)	0	4	5	9	0	0	1	2	11	8	3	5	6
I(J,Q)	14	10	9	5	14	14	13	12	3	6	11	9	8
F (J,Q)	****	71.4	64.3	64.3	****	****	92.9	85.7	78.6	57.1	78.6	64.3	57.1
F		82.4											
HYP 1=A	1	0	0	0	1	1	1	1	1	0	1	0	0
			*			*					*		

GROUP B
QUESTION 16 - CONTINUED

DOCUMENT NUMBERS

ANALYST	B693	B382	B671	B980	B920	A740	A250	A280	A540	B950	B180	A602
201	I	R	R	I	I	R	I	R	R	R	I	R
203	I	R	R	I	I	I	I	R	R	R	I	I
204	I	R	R	I	I	R	I	I	R	R	I	R
205	I	I	I	I	I	I	I	I	I	R	I	I
206	I	R	I	I	I	I	I	I	R	R	I	R
207	I	R	I	I	I	R	I	R	R	R	I	R
208	I	R	R	R	R	R	I	R	R	R	R	I
210	I	R	I	I	I	R	I	I	R	R	I	R
211	I	R	R	I	R	R	R	R	R	R	I	I
212	R	R	I	I	I	R	I	R	R	R	I	R
213	I	R	R	I	I	R	I	R	R	R	R	R
214	I	I	I	I	I	R	I	I	I	R	I	R
215	I	R	I	I	I	R	I	R	R	R	I	R
216	I	R	I	I	I	I	I	I	I	R	I	R
R(J,Q)	1	12	6	1	2	10	1	8	11	14	2	10
I(J,Q)	13	2	8	13	12	4	13	6	3	0	12	4
F (J,Q)	92.9	85.7	57.1	92.9	85.7	71.4	92.9	57.1	78.6	****	85.7	71.4
F (Q)		79.4										
F		82.4										
HYP 1=A	1	1	0	1	1	0	1	0	1	1	1	0
		*					*				*	

APPENDIX D

In order to test Hypothesis 3, the concensus of the members of each group must be observed and compared. If for a group $I_{dq} < R_{dq}$, then the members judged the document relevant, R. Likewise, if $R_{dq} < I_{dq}$, the judgment for the document was I. If $R_{dq} = I_{dq}$, then no decision as to the relevance of the document was reached and was denoted by "-". Table 4 is a listing of the judgments for all documents. The document numbers indicate the relative location within the questionnaire and are not the document numbers actually assigned to the documents.

Table 4. Results of the Relevance Assessments by Group, Question, and Document

Question	Group	Documents																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	A	I	-	I	I	I	I	I	I	I	I	I	I	I	I	I										
	B	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I										
3	A	I	I	I	I	I	I	I	I	I	I	I	I	-	I	I	I									
	B	I	I	I	I	I	R	I	R	R	I	I	I	R	I	R	I									
4	A	I	I	I	I	I	I	I	I	I	I	I	I													
	B	I	I	I	I	I	I	I	I	R	R	I	I													
5	A	I	R	R	R	R	I	R	R	I	R	R														
	B	R	R	R	R	R	R	R	R	I	R	R														
7	A	R	R	R	R	-	I	I	I	I	I	I	I	R	I	I	-	I	-	I	I	I				
	B	R	R	R	R	R	R	I	I	R	R	-	R	R	R	I	I	I	R	I	I	R				
8	A	I	I	I	R	I	R	R	I	I	I	I	I	I												
	B	I	R	I	-	I	R	R	I	I	I	I	I	I												
9	A	I	I	I	I	I	I	I	R	I	I	I	I													
	B	I	I	I	I	I	I	I	R	I	I	I	R													
10	A	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
	B	I	-	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	R	I	I	I	

Table 4. (Continued)

Question	Group	Documents																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
11	A	I	I	R	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	R	I
	B	I	I	R	I	I	I	R	I	I	I	I	-	I	I	I	I	I	I	I	I	I	I	I	R	I
13	A	-	I	I	I	R	I	I	I	I	R	I	I	I	I	I	I	-	I	I	I					
	B	I	I	I	I	R	I	I	I	I	R	I	I	I	I	I	I	R	I	I	I					
15	A	R	I	I	I	I	I	I	I	I	I															
	B	-	I	I	I	I	I	I	I	I	I															
16	A	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
	B	I	I	I	R	I	I	I	I	R	R	I	I	I	I	R	I	I	I	R	I	R	R	R	I	R

APPENDIX E

Data for testing Hypothesis 5 were selected at random from the respective agreement scores of the two analysts groups. These data were collected into the tables of this appendix with the required cumulative step functions for each sample.

Table 5. Group A Samples From Randomly Matched Documents

Sample											
1	2	3	4	5	6	7	8	9	10	11	12
100.0	100.0	100.0	92.9	100.0	100.0	92.9	100.0	100.0	78.6	100.0	57.1
78.6	85.7	100.0	78.6	100.0	100.0	92.9	71.4	92.9	100.0	64.3	100.0
92.9	100.0	100.0	92.9	92.9	92.9	100.0	92.9	100.0	100.0	92.9	100.0
100.0	100.0	57.1	78.6	100.0	100.0	100.0	100.0	100.0	85.7	85.7	100.0
100.0	92.9	100.0	100.0	100.0	92.9	85.7	85.7	100.0	92.9	100.0	85.7
92.9	100.0	92.9	100.0	92.9	78.6	71.4	92.9	100.0	100.0	100.0	92.9
57.1	78.6	50.0	100.0	78.6	78.6	78.6	78.6	92.9	92.9	100.0	92.9
100.0	71.4	78.6	92.9	100.0	100.0	100.0	71.4	100.0	85.7	100.0	85.7
71.4	92.9	92.9	100.0	71.4	92.9	100.0	100.0	100.0	100.0	78.6	100.0
100.0	92.9	50.0	92.9	100.0	92.9	57.1	57.1	92.9	100.0	92.9	100.0
100.0	100.0	100.0	92.9	78.6	100.0	92.9	92.9	100.0	100.0	100.0	100.0
64.3	50.0	100.0	100.0	100.0	100.0	100.0	100.0	85.7	92.9	100.0	85.7
64.3	100.0	78.6	92.9	85.7	78.6	100.0	100.0	85.7	100.0	92.9	92.9
92.9	100.0	100.0	78.6	92.9	100.0	100.0	100.0	100.0	100.0	92.9	64.3
100.0	78.6	92.9	100.0	100.0	85.7	100.0	92.9	100.0	100.0	100.0	92.9
100.0	100.0	100.0	71.4	71.4	64.3	100.0	100.0	92.9	78.6	100.0	85.7
50.0	100.0	100.0	100.0	100.0	100.0	85.7	71.4	100.0	71.4	92.9	64.3
92.9	100.0	100.0	85.7	100.0	100.0	85.7	85.7	71.4	100.0	100.0	57.1
92.9	100.0	78.6	78.6	64.3	92.9	92.9	100.0	100.0	50.0	100.0	100.0
92.9	92.9	92.9	92.9	85.7	57.1	92.9	92.9	100.0	50.0	85.7	71.4
85.7	100.0	100.0	100.0	100.0	92.9	57.1	100.0	92.9	100.0	71.4	71.4
50.0	71.4	100.0	57.1	92.9	100.0	100.0	100.0	100.0	100.0	78.6	92.9
57.1	100.0	100.0	64.3	100.0	100.0	100.0	64.3	92.9	100.0	100.0	92.9
50.0	64.3	57.1	100.0	92.9	71.4	100.0	100.0	71.4	92.9	100.0	78.6
57.1	100.0	92.9	100.0	100.0	57.1	71.4	100.0	64.3	85.7	71.4	100.0

Table 6. Group A Samples From Machine Matched Documents

Sample											
1	2	3	4	5	6	7	8	9	10	11	12
71.4	85.7	100.0	100.0	85.7	100.0	64.3	92.9	50.0	50.0	64.3	50.0
71.4	100.0	100.0	50.0	92.9	71.4	71.4	71.4	50.0	92.9	100.0	100.0
100.0	64.3	85.7	100.0	64.3	64.3	92.9	85.7	57.1	92.9	85.7	85.7
85.7	100.0	100.0	50.0	100.0	100.0	100.0	85.7	85.7	50.0	78.6	64.3
50.0	100.0	100.0	100.0	71.4	100.0	71.4	100.0	100.0	50.0	100.0	57.1
64.3	92.9	92.9	57.1	85.7	100.0	71.4	100.0	92.9	50.0	92.9	64.3
100.0	64.3	92.9	100.0	100.0	57.1	100.0	57.1	100.0	57.1	92.9	92.9
57.1	92.9	64.3	92.9	92.9	57.1	50.0	57.1	64.3	100.0	100.0	71.4
85.7	92.9	100.0	64.3	100.0	50.0	92.9	50.0	92.9	92.9	50.0	50.0
50.0	71.4	50.0	92.9	57.1	100.0	71.4	50.0	71.4	100.0	85.7	100.0
85.7	71.4	100.0	71.4	78.6	100.0	71.4	92.9	100.0	71.4	57.1	100.0
57.1	71.4	95.9	71.4	57.1	100.0	100.0	100.0	64.3	85.7	64.3	50.0
92.9	57.1	50.0	85.7	85.7	57.1	100.0	50.0	100.0	100.0	92.9	100.0
100.0	100.0	85.7	100.0	85.7	64.3	92.9	64.3	64.3	100.0	100.0	71.4
100.0	85.7	92.9	64.3	85.7	57.1	85.7	92.9	85.7	64.3	57.1	57.1
64.3	100.0	92.9	64.3	92.9	92.9	100.0	85.7	100.0	57.1	100.0	78.6
92.9	71.4	78.6	85.7	92.9	57.1	100.0	50.0	92.9	100.0	100.0	100.0
85.7	71.4	92.9	92.9	78.6	100.0	92.9	92.9	64.3	57.1	100.0	100.0
100.0	57.1	92.9	78.6	64.3	64.3	78.6	100.0	64.3	100.0	92.9	100.0
71.4	100.0	57.1	85.7	100.0	71.4	100.0	57.1	85.6	85.7	85.7	100.0
85.7	100.0	85.7	100.0	100.0	85.7	85.7	100.0	85.7	100.0	85.7	85.7
100.0	57.1	100.0	100.0	50.0	100.0	85.7	100.0	92.9	57.1	85.7	100.0
57.1	57.1	100.0	57.1	64.3	100.0	85.7	85.7	64.3	92.9	100.0	100.0
64.3	57.1	85.7	71.4	64.3	100.0	71.4	71.4	100.0	100.0	92.9	64.3
71.4	100.0	100.0	71.4	57.1	100.0	92.9	71.4	100.0	85.7	57.1	57.1

Table 7. Group B Samples From Randomly Matched Documents

Sample											
1	2	3	4	5	6	7	8	9	10	11	12
100.0	100.0	100.0	78.6	100.0	92.9	92.9	85.7	85.7	50.0	64.3	85.7
92.9	92.9	100.0	85.7	100.0	100.0	100.0	71.4	92.9	92.9	92.9	92.9
57.1	92.9	92.9	92.9	85.7	78.6	100.0	71.4	100.0	92.9	100.0	92.9
57.1	78.6	64.3	85.7	92.9	85.7	92.9	100.0	100.0	64.3	78.6	92.9
92.9	92.9	92.9	100.0	100.0	92.9	64.3	85.7	92.9	85.7	100.0	92.9
100.0	85.7	78.6	85.7	85.7	85.7	78.6	71.4	92.9	92.9	100.0	92.9
57.1	78.6	78.6	100.0	92.9	100.0	100.0	100.0	92.9	92.9	64.3	100.0
100.0	64.3	50.0	85.7	85.7	57.1	100.0	85.7	100.0	92.9	92.9	64.3
64.3	57.1	100.0	71.4	85.7	100.0	92.9	100.0	92.9	100.0	100.0	92.9
100.0	100.0	78.6	85.7	100.0	71.4	85.7	64.3	57.1	64.3	100.0	100.0
100.0	92.9	100.0	57.1	57.1	100.0	100.0	92.9	92.9	100.0	78.6	100.0
57.1	78.6	92.9	100.0	100.0	100.0	100.0	100.0	92.9	85.7	92.9	92.9
78.6	85.7	57.1	92.9	85.7	78.6	100.0	100.0	57.1	100.0	100.0	71.4
100.0	100.0	78.6	85.7	100.0	100.0	92.9	100.0	85.7	92.9	71.4	78.6
92.9	85.7	100.0	100.0	92.9	92.9	100.0	92.9	92.9	92.9	78.6	78.6
92.9	100.0	85.7	78.6	85.7	78.6	85.7	92.9	92.9	50.0	92.9	78.6
57.1	92.9	64.3	78.6	100.0	100.0	78.6	85.7	92.9	85.7	100.0	92.9
100.0	92.9	100.0	64.3	100.0	78.6	92.9	92.9	85.7	92.9	85.7	64.3
92.9	92.9	78.6	78.6	57.1	100.0	157.1	92.9	100.0	57.1	100.0	92.9
71.4	92.9	100.0	57.1	85.7	71.4	92.9	92.9	92.9	78.6	100.0	64.3
85.7	92.9	100.0	100.0	85.7	100.0	78.6	100.0	100.0	92.9	64.3	85.7
57.1	85.7	78.6	78.6	92.9	92.9	100.0	78.6	85.7	85.7	100.0	85.7
78.6	100.0	100.0	64.3	78.6	100.0	92.9	78.6	85.7	100.0	57.1	57.1
78.6	57.1	78.6	100.0	85.7	85.7	92.9	92.9	85.7	100.0	78.6	92.9
85.7	100.0	92.9	92.9	100.0	71.4	78.6	100.0	57.1	100.0	85.7	100.0

Table 8. Group B Samples From Machine Matched Documents

Sample											
1	2	3	4	5	6	7	8	9	10	11	12
85.7	71.4	100.0	92.9	85.7	71.4	78.6	64.3	57.1	64.3	71.4	57.1
57.1	85.7	71.4	64.3	78.6	71.4	57.1	57.1	71.4	64.3	64.3	85.7
50.0	64.3	64.3	85.7	57.1	57.1	64.3	85.7	78.6	64.3	57.1	71.4
64.3	85.7	100.0	57.1	71.4	85.7	71.4	71.4	64.3	64.3	71.4	78.6
57.1	92.9	85.7	85.7	57.1	92.9	57.1	100.0	57.1	57.1	71.4	85.7
78.6	50.0	64.3	64.3	64.3	57.1	57.1	85.7	78.6	71.4	57.1	78.6
85.7	78.6	64.3	71.4	57.1	92.9	85.7	78.6	64.3	78.6	57.1	64.3
64.3	71.4	78.6	64.3	71.4	78.6	57.1	64.3	57.1	64.3	64.3	85.7
71.4	71.4	92.9	64.3	50.0	64.3	85.7	57.1	71.4	78.6	57.1	71.4
64.3	85.7	64.3	78.6	92.9	85.7	57.1	71.4	57.1	100.0	92.9	85.7
71.4	85.7	71.4	57.1	71.4	85.7	85.7	78.6	92.9	57.1	64.3	64.3
92.9	85.7	85.7	71.4	78.6	50.0	85.7	85.7	64.3	71.4	57.1	64.3
64.3	78.6	57.1	71.4	71.4	85.7	85.7	57.1	92.9	71.4	92.9	71.4
85.7	92.9	71.4	57.1	71.4	64.3	64.3	57.1	57.1	92.9	71.4	85.7
92.9	64.3	85.7	64.3	64.3	78.6	85.7	85.7	71.4	64.3	71.4	92.9
78.6	71.4	71.4	78.6	50.0	85.7	100.0	78.6	71.4	85.7	57.1	100.0
50.0	57.1	71.4	71.4	64.3	78.6	64.3	64.3	50.0	71.4	78.6	71.4
64.3	57.1	64.3	71.4	71.4	57.1	85.7	71.4	78.6	78.6	50.0	50.0
100.0	64.3	78.6	71.4	57.1	78.6	71.4	50.0	64.3	71.4	71.4	71.4
92.9	71.4	64.3	78.6	71.4	57.1	92.9	78.6	64.3	71.4	71.4	71.4
85.7	100.0	64.3	71.4	85.7	71.4	64.3	57.1	64.3	92.9	50.0	64.3
50.0	85.7	92.9	57.1	57.1	64.3	85.7	71.4	71.4	92.9	71.4	92.9
64.3	64.3	57.1	85.7	64.3	71.4	71.4	78.6	64.3	64.3	57.1	100.0
57.1	92.9	64.3	57.1	57.1	85.7	57.1	71.4	71.4	64.3	85.7	64.3
57.1	57.1	71.4	57.1	92.9	85.7	78.6	57.1	57.1	71.4	71.4	64.3

Table 9. Cumulative Step Function and Differences for Testing Hypothesis 5

Sample	1						2						3						4					
Group	A			B			A			B			A			B			A			B		
Intervals	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D
100-96	8	6	2	7	1	6	14	8	6	6	1	5	13	9	4	9	2	7	10	7	3	6	0	6
95-91	14	8	6	12	4	8	18	11	7	15	4	11	18	16	2	13	4	9	17	10	7	9	1	8
90-86	14	8	6	12	4	8	18	11	7	15	4	11	18	16	2	13	4	9	17	10	7	9	1	8
85-81	15	13	2	14	8	6	19	13	6	19	10	9	18	20	2	14	7	7	18	13	5	15	4	11
80-76	16	13	3	17	10	7	21	13	8	22	12	10	21	21	0	21	9	12	22	14	8	20	8	12
75-71	17	17	0	18	12	6	23	18	5	22	17	5	21	21	0	21	15	6	23	18	5	21	14	7
70-66	17	17	0	18	12	6	23	18	5	22	17	5	21	21	0	21	15	6	23	18	5	21	14	7
65-61	19	20	1	19	18	1	24	20	4	23	21	2	21	22	1	23	23	0	24	21	3	23	19	4
60-56	22	23	1	25	22	3	24	25	1	25	24	1	23	23	0	24	25	1	25	23	2	25	25	0
55-50	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0

Sample	5						6						7						8					
Group	A			B			A			B			A			B			A			B		
Intervals	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D
100-96	13	5	8	9	0	9	11	12	1	10	0	10	12	7	5	9	1	8	12	6	6	8	1	7
95-91	18	9	9	13	2	11	17	13	4	14	2	12	17	12	5	17	2	15	17	10	7	15	1	14
90-86	18	9	9	13	2	11	17	13	4	14	2	12	17	12	5	17	2	15	17	10	7	15	1	14
85-81	20	14	6	22	4	18	18	14	4	17	9	12	20	16	4	19	10	9	19	14	5	19	5	14
80-76	22	16	6	23	6	17	21	14	7	21	13	7	21	17	4	23	12	11	20	14	6	21	10	11
75-71	24	17	7	23	13	10	22	16	6	24	17	7	23	23	0	23	15	8	23	17	6	24	15	9
70-66	24	17	7	23	13	10	22	16	6	24	17	7	23	23	0	23	15	8	23	17	6	24	15	9
65-61	25	21	4	23	17	6	23	19	4	24	20	4	23	24	1	24	19	5	24	18	6	25	18	7
60-56	25	24	1	25	23	2	25	24	1	25	24	1	25	24	1	25	25	0	25	21	4	25	24	1
55-50	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0

Table 9. Cumulative Step Function and Difference for Testing Hypothesis 5. (Continued)

Sample	9						10						11						12					
Group	A			B			A			B			A			B			A			B		
Intervals	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D	R	M	D
100-96	14	7	7	5	0	5	13	8	5	6	1	5	13	18	5	10	0	10	8	10	2	3	2	1
95-91	20	11	9	16	2	14	17	12	5	15	4	11	18	13	5	14	2	12	14	11	3	13	4	9
90-86	20	11	9	16	2	14	17	12	5	15	4	11	18	13	5	14	2	12	14	11	3	13	4	9
85-81	22	15	7	22	2	20	20	15	5	19	5	14	20	18	2	16	3	13	18	13	5	17	10	7
80-76	22	15	7	22	5	17	22	15	7	20	8	12	22	19	3	20	4	16	19	14	5	20	12	8
75-71	24	16	8	22	11	11	23	16	7	20	14	6	24	19	5	21	13	8	21	16	5	21	17	4
70-66	24	16	8	22	11	11	23	16	7	20	14	6	24	19	5	21	13	8	21	16	5	21	17	4
65-61	25	22	3	22	18	4	23	17	6	22	22	0	25	21	4	24	16	8	23	19	4	24	23	1
60-56	25	23	2	25	24	1	23	21	2	23	24	1	25	24	1	25	23	2	25	22	3	25	24	1
55-50	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0	25	25	0

APPENDIX F

The frequencies of the R judgments were found and recorded in one of the following four tables. For the randomly selected documents there were 202 judgments, for the machine matched documents 45 judgments. The relative frequency and the cumulative relative frequency were also tabulated for each R judgment and recorded in the following tables.

Table 10. Frequency, Relative, and Cumulative Relative Frequency of R Judgments of All Documents for Group A

R-Judgments	Frequency	Relative Frequency	Cumulative Frequency
I	148	.733	.733
4	9	.045	.778
5	12	.059	.837
6	7	.035	.872
7	7	.035	.907
8	3	.015	.922
9	1	.005	.927
10	4	.020	.947
R	11	.053	1.000

Table 11. Frequency, Relative and Cumulative Relative Frequency of R Judgments of Machine Matched Documents for Group A

R-Judgments	Frequency	Relative Frequency	Cumulative Frequency
I	22	.488	.488
4	3	.067	.555
5	2	.045	.600
6	3	.067	.667
7	3	.067	.734
8	2	.045	.779
9	1	.021	.800
10	2	.045	.845
R	7	.155	1.000

Table 12. Frequency, Relative and Cumulative Relative Frequency of R Judgments of All Documents for Group B

R-Judgments	Frequency	Relative Frequency	Cumulative Frequency
I	117	.579	.579
4	9	.045	.624
5	14	.069	.693
6	7	.035	.728
7	5	.025	.753
8	12	.059	.812
9	9	.045	.857
10	11	.054	.911
R	18	.089	1.000

Table 13. Frequency, Relative and Cumulative Relative Frequency of R Judgments of Machine Matched Documents for Group B

R-Judgments	Frequency	Relative Frequency	Cumulative Frequency
I	5	.111	.111
4	4	.089	.200
5	7	.155	.355
6	2	.045	.400
7	2	.045	.445
8	4	.089	.534
9	3	.067	.601
10	7	.155	.756
R	11	.244	1.000

APPENDIX G

This appendix consists of two tables, one for each group, recording the times in minutes for the analysts to judge the relevance of all documents of a question to the question. The "x" notation implies that the analysts failed to record his time on the questionnaire. The total time, the average time per question and the average time per document are all recorded in minutes.

Table 14. Judgment Times for Group A

Analyst Numbers	Question Numbers 2	3	4	5	7	8	9	10	11	13	15	16
101	x	x	20	10	20	x	15	25	x	x	15	x
102	8	5	5	4	9	5	4	12	5	6	3	9
103	10	10	10	8	10	10	8	15	10	10	8	10
104	4	9	4	5	10	5	4	9	11	8	4	9
105	7	8	5	5	9	6	6	8	7	7	3	8
106	2	3	2	3	3	2	3	4	4	3	1	4
107	6	6	5	3	5	4	3	5	7	5	3	8
108	4	4	2	2	3	3	2	2	3	2	1	2
109	10	8	3	5	10	7	5	10	8	x	3	10
110	7	10	5	8	10	9	5	6	11	6	5	7
111	14	11	7	8	12	15	8	11	11	15	4	16
112	4	6	5	3	5	4	4	5	5	6	3	7
113	30	20	15	10	10	15	12	15	20	15	8	13
114	3	5	3	3	3	x	3	5	3	3	3	3
Total Time/ Question	109	105	91	77	119	85	82	132	105	86	64	106
Average Time/ Question	8.4	8.1	6.5	5.5	8.5	7.1	5.9	9.4	8.1	7.2	4.6	8.2
Average Time/ Document	.56	.51	.54	.50	.40	.55	.49	.41	.34	.36	.46	.33

Table 15. Judgment Times for Group B

Analyst Number	Question Numbers 2	3	4	5	7	8	9	10	11	13	15	16
201	15	12	15	10	15	15	15	15	25	20	8	25
203	10	15	10	15	20	10	15	25	20	15	10	20
204	5	4	3	4	5	6	4	6	7	5	3	7
205	8	7	6	5	7	8	4	x	5	5	2	6
206	15	13	12	6	15	7	7	20	10	13	4	17
207	60	15	25	20	35	20	20	35	35	25	15	45
208	8	7	6	5	9	5	9	9	8	11	7	9
210	9	6	5	4	12	5	4	11	7	6	2	10
211	8	10	11	10	20	9	9	15	12	8	8	10
212	17	11	10	7	13	8	8	18	11	10	10	23
213	15	8	10	5	10	8	10	20	15	15	6	15
214	1	1	1	1	2	1	1	1	1	1	1	1
215	10	12	8	7	15	11	6	14	13	13	4	18
216	10	7	10	7	10	10	3	10	7	10	5	7
Total Time/ Question	191	128	132	106	188	123	115	199	176	157	85	213
Average Time/ Question	13.5	9.2	9.4	7.6	13.4	8.8	8.2	15.3	12.6	11.2	6.1	15.2
Average Time/ Document	.91	.57	.79	.69	.64	.68	.69	.66	.52	.56	.61	.61

BIBLIOGRAPHY

Barhydt, Gordon C., "A Comparison of Relevance Assessments by Three Types of Evaluator," Proceedings of the American Documentation Institute, I(1964), 383-385.

Bornstein, Harry, "A Paradigm for a Retrieval Effectiveness Experiment," American Documentation, XII (October 1961), 254-259.

Cleverdon, Cyril W., "Testing of Index Language Devices," ASLIB Proceedings, XV (April 1963), 107.

Cleverdon, Cyril W., ASLIB Cranfield Research Project, Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield: ASLIB, 1962.

Fels, E. M., "Evaluation of the Performance of an Information-Retrieval System by Modified Mooers Plan," American Documentation, XIV (January 1963), 28-34.

Hillman, Donald J., "The Notion of Relevance(I)," American Documentation, XV(January 1964), 29.

Klempner, Irving M., "Methodology for the Comparative Analysis of Information Storage and Retrieval Systems: A Critical Review," American Documentation, XV(July 1964), 213.

Lancaster, F. W. and J. Mills, "Testing Indexes and Index Devices: The ASLIB Cranfield Project," American Documentation, XV(January 1964), 4-12.

O'Connor, John, "Reviews," Journal of Documentation, XVII(December, 1961), 259-260.

Optner, Stanford L., Systems Analysis for Business and Industrial Problem Solving. Englewood Cliffs: Prentice-Hall, 1965.

Rath, G. J., A. Resnick, and T. R. Savage, "Comparison of Four Types of Lexical Indicators of Content," American Documentation, XII (April 1961), 126-130.

Rath, G. J., A. Resnick, and T. R. Savage, "The Formation of Abstracts by the Selection of Sentences," American Documentation, XII (April 1961), 139-143.

Rees, Alan M., "Relevancy and Pertinency in Indexing," American Documentation, XIII(January 1962), 23.

Resnick, A., "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents," Science, CXXXIV (October 1961), 1004-1005.

Sayer, John, "Do Present Information Services Serve the Engineer," Data Processing Magazine, VII(February 1965), 24-25, 64-65.

Siegel, Sidney, Nonparametric Statistics for the Behavioral Sciences. New York: McGraw Hill, 1956.

Summary of the Study Conference in Evaluation of Document Searching Systems and Procedures. Washington: National Science Foundation, 1965. (Multilithed).

Swanson, Don R., "Information Retrieval: State of the Art," Proceeding of the Western Joint Computer Conference, XIX(1961), 242-243.

Swanson, Don R., "Searching Natural Language Text by Computer," Science, CXXXII(October 1960), 1099-1104.

Turanski, William J., UNIVAC Data Automated Systems: Basic Concepts of Information Retrieval. AD 149 510. Remington Rand UNIVAC, 1958.